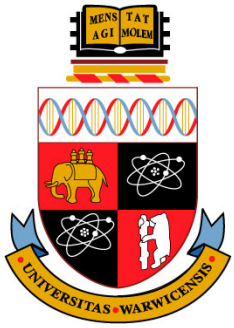


University of Warwick

MOAC Thesis

Functional and Structural Insights into MmyJ, An ArsR-Like Transcriptional Repressor

Author:

Matthew James Lougher

0700304

Supervisors:

Christophe Corre

Józef Lewandowski

Daniel Mitchell

September 25, 2015

Contents

Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosphy. It has been composed by myself and has not been submitted in any previous application for any degree.

The work presented (including data generated and data analysis) was carried out by the author except in the cases outlined below:

- HADDOCK Model in Chapter 2 was generated by Alex Fulwood
- MmyJ cloning and pET151/D-TOPO plasmid preparation in Chapter 3 was carried out by Christophe Corre
- Mass Spectrometry data in Chapter 3 was obtained by Lijiang Song
- AUC data in Chapter 4 was obtained by staff at Birmingham Biophysical Characterisation Facility
- NMR data in Chapter 6 was acquired in collaboration with Ivan Prokes, Józef Lewandowski and Daniel Griffiths (solution state) and Józef Lewandowski and Carl Öster (solid state)
- X-Ray Diffraction data in Chapter 6 was collected by Dean Rea and processed by Vilmos Fulop

Matthew James Lougher

Acknowledgements

First and foremost, I would like to acknowledge and thank my three supervisors: Christophe Corre, Józef Lewandowski and Dan Mitchell. Needless to say, without their support and guidance this project would not have been completed, even though it ended up going in a different direction to our original plan. I would also like to thank my advisory panel, consisting of Vilmos Fulop, Claudia Blindauer and Stephen Brown, whose insights and suggestions often helped me throughout the project.

For day to day help and support in the lab I would like to extend my thanks to the Corre and Challis groups within the Chemical Biology Research Facility at the University of Warwick, as well as Anne Smith for maintaining the lab. Individually, I'd like to thank John Sidda for helping me find my feet and showing me the basics in the lab (very different to the physics labs I had previously worked in). I'd also like to thank Vincent Poon for helping me with the bioinformatics and RT-PCR, Alex Fulwood for helping with the creation of the HADDOCK model and Gideon Idowu and Shanshan Zhou for supplying the methylenomycin and methylenomycin furan compounds respectively, without which I would not have been able to complete my assays. I'd also like to thank Zdenek Kamenik for teaching me how to use the FPLC system, and Dean Rea for helping set up crystal trials. I'd like to thank Vilmos for analysing the resulting crystal data, as well as Lijiang Song for running mass spectrometry samples for me. For their help in running NMR experiments, I'd also like to thank Daniel Griffiths, Carl Öster, Józef Lewandowski and Ivan Prokes. Finally I'd like to extend my thanks to everyone in MOAC for helping me over the last 4 years, as well as the EPSRC for providing funding, allowing me to take the opportunity to work on this project.

On a personal note I'd like to thank my friends and family for their help and support during the time I have been performing this work. I'd like to thank my parents Beverley and John for supporting me in my decision to pursue a PhD, as well as my brother Alexander. I'd also like to thank Alison for keeping me motivated and spending many evenings helping me proof read this work.

List of Figures

List of Tables

Abbreviations Used

AUC	Analytical Ultracentrifugation
bp	Base Pair
BLAST	Basic Local Alignment Search Tool
CD	Circular Dichroism
cDNA	Complementary DNA
CDS	Coding Sequence
CFE	Cell Free Extract
CP	Cross Polarisation
DMSO	Dimethyl Sulfoxide
DTT	Dithiothreitol
DSS	4,4-Dimethyl-4-Silapentane-1-Sulfonic Acid
EMSA	Electrophoretic Mobility Shift Assay
FID	Free Induction Decay
FPLC	Fast Protein Liquid Chromatography
HPLC	High Pressure Liquid Chromatography
HSQC	Heteronuclear Single Quantum Coherence
HTH	Helix Turn Helix
IMAC	Immobilised Metal Affinity Chromatography
IPTG	Isopropyl β-D-1-Thiogalactopyranoside
lpl	Left Polarised Light
MAS	Magic Angle Spinning
Mm	Methylenomycin
MPD	2-Methyl-2,4-pentanediol
MS	Mass Spectrometry
NMR	Nuclear Magnetic Resonance
NTA	Nitrilotriacetic Acid
OD	Optical Density
PDB	Protein Data Bank
PCR	Polymerase Chain Reaction
PEG	Polyethylene Glycol
ppm	Parts Per Million
RMSD	Root Mean Squared Deviation
rpl	Right Polarised Light
RT-PCR	Reverse Transcription Polymerase Chain Reaction
SD8	Simocyclinone D8
SDS-PAGE	Sodium Dodecyl Sulphate Poly-Acrylamide Gel Electrophoresis
Se-Met	Selenomethionine
SEC	Size Exclusion Chromatography
sGFP	Superfolded Green Fluorescent Protein
SPR	Surface Plasmon Resonance
TEV	Tobacco Etch Virus
TCEP	Tris(2-carboxyethyl)phosphine
TMS	Tetramethylsilane
TROSY	Transverse Relaxation-Optimised Spectroscopy
XRD	X-Ray Diffraction

Abstract

MmyJ is a protein encoded in the methylenomycin antibiotic gene cluster in *Streptomyces coelicolor* A3(2). It was identified as a novel member of the ArsR family of transcription repressors. In depth bioinformatic analyses were carried out to compare this to other members of the ArsR family, leading to high confidence of its classification. An expression system was engineered to investigate MmyJ *in vitro*, leading to the discovery of the presence of a covalently bonded dimer, which is unusual for ArsR family proteins and as such was thought to be an artefact of purification. A C49S mutant was then also engineered without the capacity to form covalent dimers, such that the two variants could be investigated side by side. Work was carried out to investigate the stability of MmyJ, and it was found that its secondary structure denatured above temperatures of 40°C and that the protein is only robust to a single freeze/thaw cycle. Electrophoretic mobility shift assays were then used to prove that MmyJ binds specifically to a 13-1-13 semi conserved inverted repeat overlapping the -35 region of both *mmyJ* and *mmr* promoters. This indicates that MmyJ not only regulates its own expression, but also that of Mmr, an efflux pump that removes methylenomycin A from the cell upon biosynthesis. It was also demonstrated that methylenomycin A caused complete dissociation of the MmyJ:DNA complex when present in a 20 times molar excess, hence suggesting that the methylenomycin A resistance mechanism is triggered in the presence of methylenomycin A. This is the first reported instance of an ArsR family protein sensing a non-metallic ligand. Structural insights into MmyJ were also sought, with a homology model produced by Phyre2 analysis complimented by circular dichroism analysis of recombinant MmyJ. X-ray diffraction data were obtained to a resolution of 2.1 Å over 300°, but the phase of these data has yet to be determined, so the crystal structure has not yet been solved.

1 Introduction

1.1 Microbial Transcription Factors

1.1.1 Transcriptional Regulation

The ability to regulate the transcription of genes is vital to bacteria; organisms must adapt to local environmental conditions in a way that allows them to survive whilst also inhibiting the growth of competing species. This has driven the evolution of a myriad of systems, including ones to take advantage of a nutrient-rich environment as well as ones that respond to extra-cellular threats via the production of secondary metabolites. However, it would be inefficient and very wasteful if all of these genes were expressed at once and would likely result in multiple systems within the cells actively competing against each other; for example, proteins designed to import specific ions into the cell may end up operating alongside other proteins exporting the same ions back out again. Hence, these systems must be regulated in real time in accordance with the changing needs of the organism.

Some systems use post-transcriptional modification as a form of regulation, inhibiting the processes between transcription of mRNA from the genomic DNA and translation into proteins [1]. However, most regulation in bacteria occurs prior to this at the transcriptional level, with systems evolving to inhibit transcription of genes into mRNA until such a time that the eventual products of these genes are required [2]. One such method of transcriptional regulation is through the presence of transcription factors: proteins which bind to a specific DNA sequence near the promotor region for the gene or operon to be regulated and either activate or repress the recruitment of RNA polymerase, hence regulating the level of transcription [3]. The amount of regulation can then be altered as cellular conditions change, triggering the release of the DNA by the transcription factor as conditions within the cell vary.

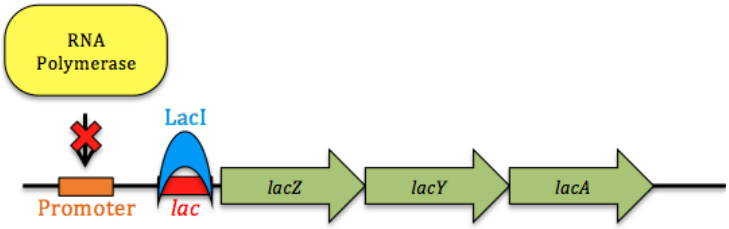
Once a transcription factor:DNA complex has been formed, either to promote or inhibit gene expression, a mechanism for dissociating this complex as required must be present. This allows alteration of gene expression levels to match the requirements of the cell, depending on the environment it is in. There are two main ways of doing this: either by having a second regulator system that controls expression of the transcription factor, or by activating/deactivating

transcription factors already present in the cell as required. In other words, either the synthesis or activity of the transcription factors can be regulated [4]. In the second case, it is common for the transcription factor:DNA complex to be stabilised or disrupted by binding to a cognate ligand. An example of this kind of transcription factor is the *Escherichia coli* protein Fur which, in the presence of Fe^{2+}, forms a complex with these ions that then associates to a specific region of DNA upstream of the genes responsible for the uptake of iron into the cell. This prevents RNA polymerase from binding and hence represses the expression of this system. Likewise, when cellular concentrations of iron fall, the Fe^{2+} ligand dissociates from Fur, which in turn dissociates from the DNA, allowing the expression of the iron uptake system in order to raise cellular concentrations of Fe^{2+} to the required level [5].

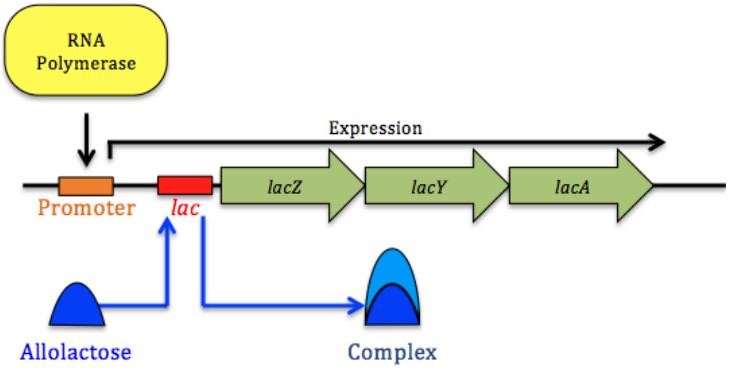
A second example of regulation of transcription factors by a binding ligand is ArsR, which regulates the expression of the gene *arsA*, encoding for an oxyanion efflux pump found on the R773 plasmid of *Escherichia coli* [6]. Unlike Fur, which is classed as a co-repressor due to the requirement of a binding ligand to initiate repression, ArsR is a transcriptional repressor which naturally forms a DNA complex until disrupted by an associating ligand. Once expressed, it binds to the promoter region of *arsA*, thus preventing RNA polymerase from transcribing the *arsA* gene. ArsR is responsive to arsenite and antimonite ions, which, upon binding, change the conformation of the transcription factor such that it dissociates from the DNA, allowing transcription of *arsA* and controlling the amount of arsenite and antimonite within the cell [7].

1.1.2 Inducible Expression Systems

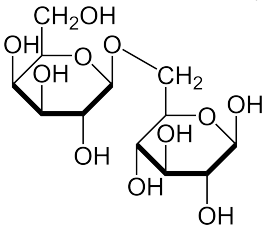
The ability of transcriptional factors to regulate gene expression has led to their development into genetically engineered inducible expression systems used to control the expression of desired proteins in laboratory conditions. One common example of this is the lac system, originally identified as regulating lactose transport and metabolism in *E. coli* strain K-12 [8]. This regulatory system is responsive to, and induced by, allolactose, a metabolite of lactose, triggering the release of LacI, a transcriptional repressor, from the *lac* operon, as illustrated in Figure 1.1 [9]. In this way, the lac system can be incorporated into engineered plasmids in order to overproduce a desired protein, with its associated gene replacing the genes downstream of the *lac* operator such that expression is inhibited. IPTG (Isopropyl β-D-1-thiogalactopyranoside),



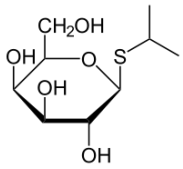
(a) Repression



(b) Induction



(c) Allolactose



(d) IPTG

Figure 1.1: Illustration of transcriptional repression by LacI. (a) Expression of *lacZ*, *lacY* and *lacA* genes is repressed due to inhibition of RNA polymerase by LacI:*lac* complex. (b) Upon formation, the LacI:allolactose complex dissociates from *lac*, allowing recruitment of RNA polymerase and expression of downstream genes. (c) and (d) show structures of allolactose and IPTG respectively.

which is a functional analogue of allolactose, also binds to LacI [10] and can be added to cell cultures in order to trigger expression of the desired gene and overproduction of the protein of interest as it readily enters the cells when present at concentrations of the order of 1 µM [11]. Extra regulation of this expression system may be added by the incorporation of RNA polymerase from T7 bacteriophage regulated by an additional *lac* operon, giving a double tier regulation of the expression system by controlling the expression of RNA polymerase as well as inhibiting the polymerase binding site [12, 13]. This system is explored further in Section 3.1.1.

1.1.3 The Helix-Turn-Helix Superfamily

Transcription factors can be divided into several structural superfamilies. One such superfamily is defined by a characteristic Helix-Turn-Helix (HTH) motif, from which the family takes its

name. This superfamily was initially defined after it was noted that several regulatory proteins, namely CAP [14] and Cro [15, 16] from *E. coli* and λ repressor [17] from bacteriophage lambda, shared a similar amino acid sequence of between 20 and 25 residues that was found to be required for DNA binding. It was observed that these three proteins have a common tertiary structure in the DNA bound state, with one α-helix fitting completely into the major groove of the binding site while the preceding helix (nearer the N terminus) lies near the DNA backbone above the groove [18]. The α-helix in the groove is commonly classified as the recognition helix [19], with the other helix often thought to stabilise the bond to the DNA, although it can also play a role in recognition [19].

From the initial recognition of three proteins with this feature, it was noted that the relative orientation of the two helices is strongly conserved [20], such that between CAP and Cro the average deviation of the 24 overlapping backbone carbons was found to be just 1.1 Å [21]. Along with the similarity of the amino acid sequences themselves, this led to the hypothesis that these three proteins were part of a much larger family of transcriptional regulators with similar characteristics, recognising the Helix-Turn-Helix as a recurring motif with two helices linked by three amino acids [22].

Since this initial classification, the HTH superfamily has grown considerably. It is defined in the NCBI database [23] as "*A large family of mostly alpha-helical protein domains with a characteristic fold; most members function as sequence-specific DNA binding domains, such as in transcription regulators. This superfamily also includes the winged helix-turn-helix domains.*" [24]. A simple search of the database on 27th August 2015 for proteins with the term "helix turn helix DNA-binding" in their name or description returned over 4.4 million hits, mostly of bacterial origin but also from archaea, eukaryotes and viruses, demonstrating how many proteins have been classified as being part of this superfamily over the intervening 30 years since the motif was initially recognised. The HTH superfamily has also been divided into many subfamilies with more defined and conserved structures and functions. Several examples of such families found in prokaryotes are detailed in Table 1.1 [25, 26].

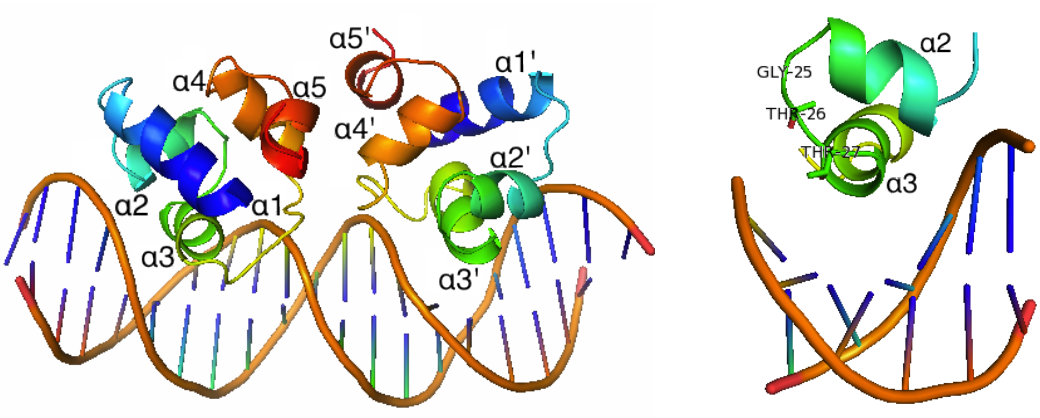
Figure 1.2 shows an example of an HTH transcription factor bound to DNA, visualised using Pymol [42]. The structure displayed is the bacteriophage 434 repressor R1-69 bound to

Family	Action	HTH Location	Function Regulated
AraC [27]	Activator	C-terminal	Sugar metabolism
ArsR [28]	Repressor	Central	Metal homeostasis
AsnC [29]	Activator/Repressor	N-terminal	Amino acid metabolism
Crp [30]	Activator/Repressor	C-terminal	Cellular processes
DeoR [31]	Repressor	N-terminal	Sugar metabolism
GntR [32]	Repressor	N-terminal	Cellular processes
IclR [33]	Repressor/Activator	N-terminal	Carbon metabolism
LacI [34]	Repressor	N-terminal	Carbon metabolism
LuxR [35]	Activator	C-terminal	Quorum sensing
LysR [36]	Activator/Repressor	N-terminal	Cellular processes
MarR [37]	Activator/Repressor	Central	Antibiotic resistance
MerR [38]	Repressor	N-terminal	Metal homeostasis
NtrC [39]	Activator	C-terminal	Nitrogen assimilation
OmpR [40]	Activator	C-terminal	Metal homeostasis
TetR [25]	Repressor	C-terminal	Antibiotic resistance

Table 1.1: Examples of families of prokaryotic proteins belonging to the HTH superfamily. In each case, the typical action is listed (where both activator and repressor are listed, the first is the primary function of the family), along with the location of the HTH motif within the peptide chain. The function that is typically regulated by the family is also listed.

a 20 base pair DNA fragment using data from [41], entry number 1RPE on the PDB [43]. It can be seen that it forms a homodimer, with each monomer interacting with the DNA strand via helices α2 and α3, with α3 being the recognition helix embedded within the major groove of the DNA sequence. Also, in accordance with observations of [22], the turn between α2 and α3 is comprised of three amino acids, here Gly25, Thr26 and Thr27, allowing for the almost perpendicular arrangement of the two binding helices, as shown explicitly in Figure 1.2b.

An example of how the interaction of a HTH transcription factor with a ligand can alter the protein's structure, disrupting DNA binding, is shown in Figure 1.3. In this instance, the



(a) (b)

Figure 1.2: Representation of the HTH transcription factor R1-69 dimer bound to a 20 base pair DNA fragment (PDB entry 1RPE, data from [41]). (a) Shows complete dimeric complex with both dimers coloured blue to red from N terminal to C terminal. (b) Shows just the HTH domain and DNA major groove, with the three residues comprising the turn labelled explicitly.

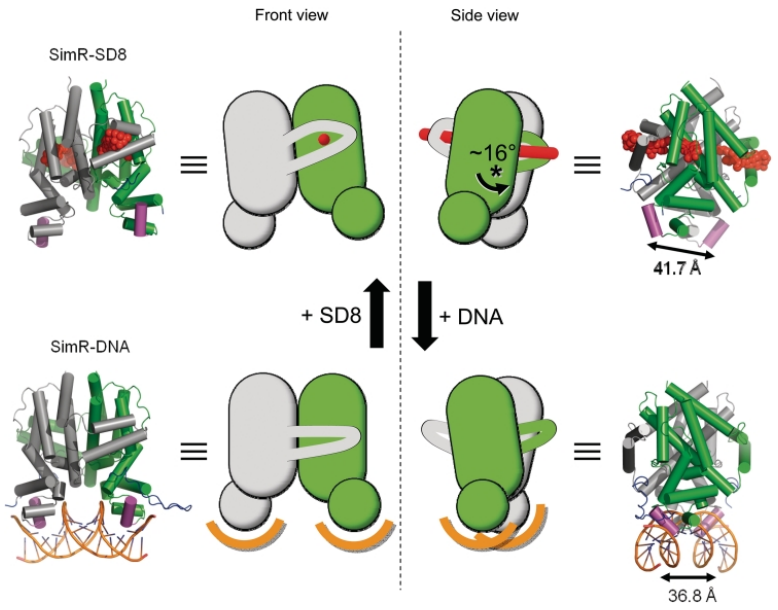


Figure 1.3: Crystal stucture of SimR in complex with both SD8 and DNA, viewed from two different directions. Two monomers of SimR are present, coloured grey and green, with the recognition helix of the HTH DNA binding domain coloured purple on both monomers. SD8 is shown in red and DNA in orange. Taken from [45].

crystal structure of SimR, a TetR family transcriptional repressor produced by *Streptomyces antibioticus*, is shown bound to both Simocyclinone D8 (SD8) and DNA. The TetR family of transcriptional repressors is defined by a 47 residue domain, containing a HTH motif and surrounding residues, which is highly conserved across TetR, QacR, CprB and EthR, both in sequence and three dimensional structure. This has allowed the identification of over 2,000 members of this family through the screening of protein and genomic databases [25]. In the system illustrated, the SD8 ligand is a DNA gyrase inhibitor native to *S. antibioticus* which is exported from the cell by SimX, the expression of which SimR regulates. Hence this is an example of a self-resistance mechanism to SD8; a secondary metabolite with antibiotic properties [44]. It can clearly be seen that the distance between the recognition helices of the two monomers, coloured purple in Figure 1.3, is altered by 4.9 Å upon receiving the SD8 ligand. This conformational change is enough to disrupt the binding of SimR to the DNA, and thus it is released, allowing transcription of SimX and export of SD8 from the cell.

1.2 The ArsR Family

1.2.1 Brief History

As previously described, ArsR is a arsenite and antimonite sensing transcriptional repressor found on *E. coli* plasmid R773 [46] that regulates the expression of *arsA* [7]. Whilst investigating similarities between ArsA and other efflux pump systems, it was noted that ArsR was "weakly but significantly similar to CadC" [47], a transcription factor that is part of a regulation system for Cd^{2+} ions located on *Staphylococcus aureus* plasmid pI258 [48, 49], hinting at the classification of a new family of metal sensing transcription factors. Interestingly, this plasmid also contained an ortholologue of the ArsA system (including ArsR) [47].

The later identification of SmtB, a Zn binding transcription factor encoded in *Synechococcus* PCC7942, as being structurally similar to ArsR [52, 53], added another protein to the classification. This group was then named the ArsR family of metalloregulatory proteins, defined as containing a highly conserved ELCVCDL region proposed to be the metal binding domain [54]. It is interesting to note that ArsR was not initially thought to have a canonical HTH DNA binding domain [47], although further investigation led to its putative identification [55]. This was later refined to exclude the adjacent cysteines, which were identified as part of the previously described metal binding site [54].

Since the definition of the ArsR family (sometimes called the ArsR-SmtB family), many more proteins have been added to the classification. In 2006 there were 1024 genetic sequences encoding for ArsR family proteins identified in the Pfam database [50, 56], with a range of

Site	Motif Pattern and Location	Classifier
α3	One or more of (i) CXCXXC, (ii) CXC or (iii) CXXD in α3	$ArsR^1$
α3N	As α3 + potential N-terminal ligands	CadC
α3N-2	As α3 + C between α2 & α3 helices + N-terminal Cys	$ArsR^2$
α4C	CXXXC in α4 + C-terminal ligands	CmtR
α5	DXHX(10)HXX(E/H) in α5	SmtB
α5C	As α5 + C-terminal His	NmtR
α5-3	As α3 + HX(6)DX(5)EHX(7)HH spanning α5 & C-terminal region	KmtR
α5-4	CC at α5 + C-terminal Cys	$ArsR^3$

Table 1.2: Description of 8 distinct metal sensing sites found in ArsR family transcription factors. Helices (α2-5) are numbered according to those in SmtB (see Section 1.2.5). Classifiers are proteins from which the motif is defined, with other candidate proteins mapped against these for assignment into one of the listed categories [50, 51]. $ArsR^{1,2,3}$ refer to ArsR orthologues from *E. coli*, *Corynebacterium glutamicum* and *Acidithiobacillus ferrooxidans* respectively.

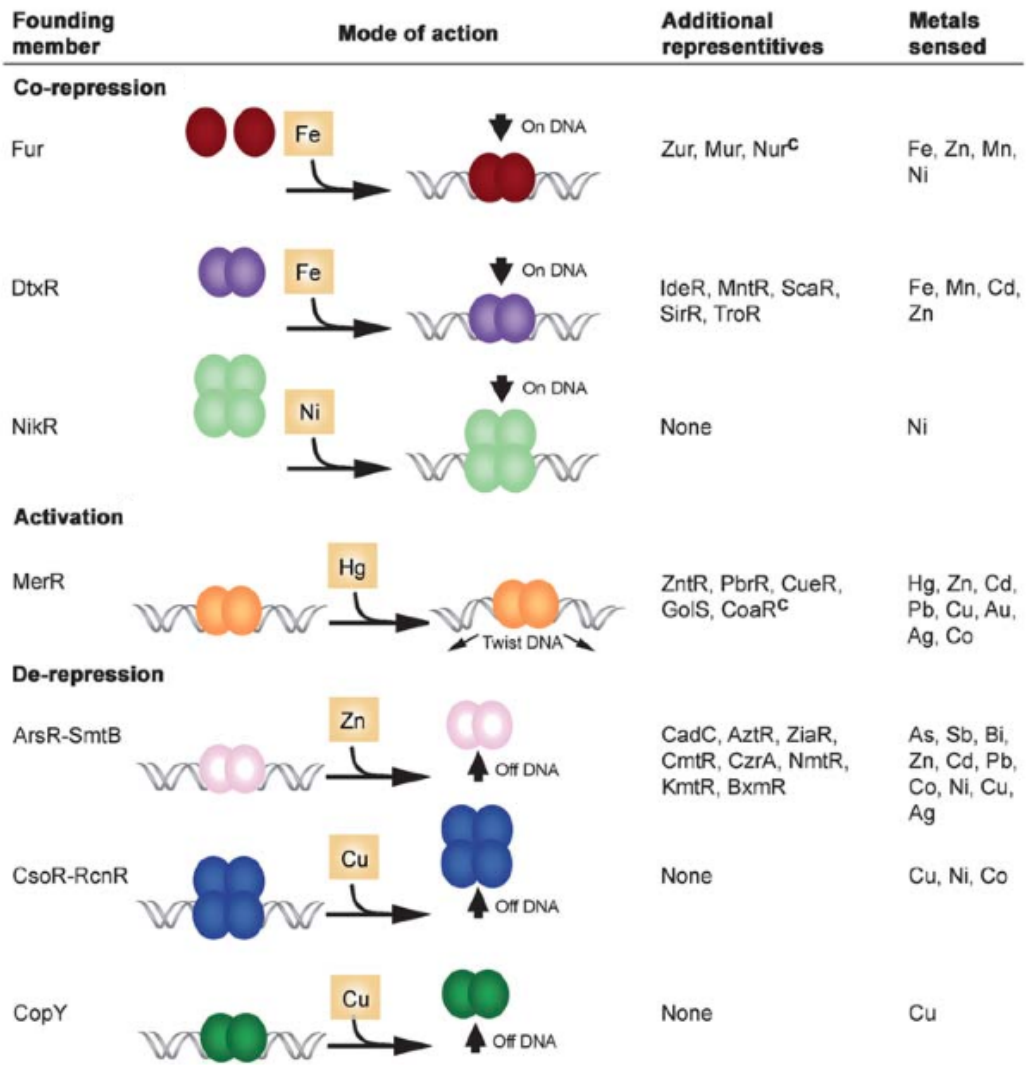


Figure 1.4: Summary of the seven major classes of soluble metal sensing transcription factors found in bacteria. In each case the founding member of the family is shown as well as a schematic of how the family typically interacts with DNA. Other example members are also shown along with a list of metals known to be sensed by different members of each family. Note: Fur and NikR members can act as DNA-binding activators when bound to iron and nickel respectively, and MerR family members can cause repression when unbound. Taken from [51].

classifying features allowing the family to be broken down into 8 smaller sub-groups containing different sensory motifs [50]. These sub groups are detailed in Table 1.2 [57, 51]. The defining factors of the ArsR family have been expanded since the initial definition of the conserved ELCVCDL motif [54], although cysteines are still the key residues in most motifs. This expansion has led to the classification of the ArsR family as one of the seven major structural bacterial families of soluble metal sensing proteins, which are summarised in Figure 1.4 [51]. Of the other families, Fur has already been described as a ligand sensing repressor, regulating the uptake of iron into *E. coli*; likewise DxtR and NikR perform similar functions. Also, the CsoR

and CopY families are similar in function to ArsR, although these families are more regimented, consisting of orthologues of only a few individual proteins. Of the seven families shown, MerR is the only one which typically remains bound to DNA regardless of ligand binding, however it still undergoes a conformational change. In this case, the DNA strand changes conformation with the transcription factor and is twisted, which allows RNA polymerase access to the regulated promoter. There are also examples of Mer family proteins, such as SoxR, acting as oxidation switches and changing conformation if oxidised, with no ligand interaction required [58].

1.2.2 Dimerisation of ArsR

It was concluded early into the investigation of ArsR that it functions as a dimer, as determined by carrying out mobility shift assays of wild type mixed with chimeric forms of ArsR. These chimeras were made by fusing a large protein, in this case BlaM, to residues at codon 91 and 143, giving the ArsR-BlaM50 and ArsR-BlaM102 chimeras respectively. In experiments with the 102 variant, three bands appeared on mobility shift assays (see Section 5.1.3): the native dimer, the chimeric dimer and the native/chimeric dimer, as shown in Figure 1.5 [59]. The same could not be said for the chimeric ArsR-BlaM50 protein, which did not exhibit dimerisation with

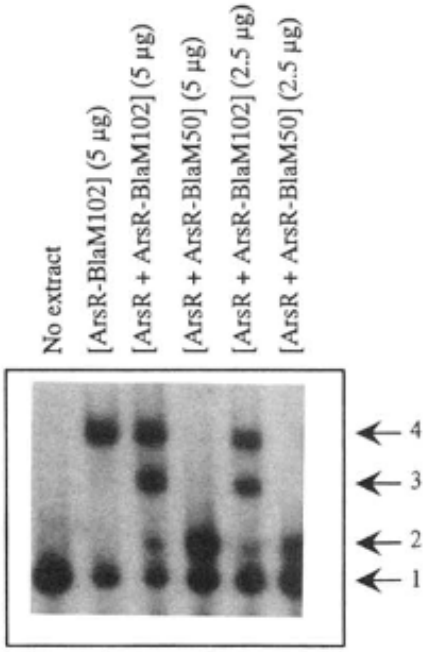


Figure 1.5: Evidence of ArsR dimerisation. Mobility Shift Assay with bands evident of four differently shifted DNA fragments being present. Band (1) corresponds to free DNA, whereas bands (2)-(4) correspond to DNA in complex with (2) ArsR dimer, (3) ArsR/ArsR-BlaM102 dimer and (4) ArsR-BlaM102 dimer. Taken from [59].

the wild type form of ArsR. This is taken to be indicative that either the DNA binding domain or dimer interface of ArsR include at least the amino acids between codon 91 and 143, as these are not present in ArsR-Blam50. For these assays, the DNA used was the 153 bp *EcoRI-DraI* fragment of plasmid pJHW1 containing the *ars* operon (this plasmid was engineered to include the *arsR* gene and *ars* operon from R773 [7]), and so these data also show that the ArsR DNA binding site lies near to the *ars* operon.

The conclusion that ArsR forms homodimers was later confirmed by the observation of an elution peak corresponding to a mass of 26 kDa, double the ArsR monomer mass of 13 kDa [7], when purified by size exclusion chromatography (see Section 4.2.1) [60]. Later work revealed that this homodimerisation was indeed essential for the function of ArsR after truncated mutations were created and mixed with full length ArsR. In order to ensure that the truncation process itself was not responsible for any alteration in DNA association, the DNA binding domain from GAL4 was fused to the truncated forms of ArsR. This assay proved that dimerisation was required for DNA binding, as well as indicating that the dimer interface lay somewhere between residues 9 and 89 of the 117 amino acid monomer [61].

1.2.3 Interactions with DNA

A 15 base pair (bp) region of DNA on the R773 plasmid, close to the *arsR* operon, was identified as being protected by DNase I footprinting and was found to contain an imperfect symmetrical dyad within its sequence [59]. As this protected region spans the -64 to -40 bases with respect to the transcriptional origin of *arsR*, and RNA polymerase is known to occupy bases from -50 to +22 [62], it was concluded that this protected region was the ArsR binding site. However, it was thought that the dimer formed by ArsR did not directly interact with the dyad [60], despite it being a typical target for DNA binding proteins [63, 64]. This was the result of high resolution techniques, which identified ArsR as binding specifically to just two sets of 4 bases: TCAT and TTTG, separated by 7 bp; a result that was likewise found on a homologous chromosomal *arsR* system in *E. coli* [60]. However, it was found that another homologue of ArsR from *Staphylococcus xylosus* did not contain this sequence, instead containing five protected regions of DNA, two of which contained dyad symmetry directly overlapping the *arsR* promoter site in this case [65].

Further study with SmtB found that this protein interacts with four distinct sites around the *smt* operator: two forming an imperfect 6-2-6 inverted repeating DNA sequence (dyad) and the other containing a 7-2-7 dyad 10 bps distant from the first pair [66]. Another difference is that while it had been shown that ArsR bound to the same DNA strand [59], SmtB was found to bind to both strands of the helix, with one part of each dyad appearing on each strand [66]. It was hypothesised that SmtB formed two distinct dimers and bound to both pairs of inverted repeats located around the *smtA* operator [66], with later evidence indicating that this may in fact be part of a monomer-dimer-tetramer system [67]. It was also found that CzrA, a zinc- and cobalt-sensing ArsR family protein, forms multimeric complexes as well, with four dimers bound across a 48 bp region near the *czr* operator containing multiple TGAA sequences. [68]. CadC was also found to bind as two distinct dimers to two sites around the *cad* operator, although binding to one of the sites was dependant on the concentration of NaCl present [69, 70].

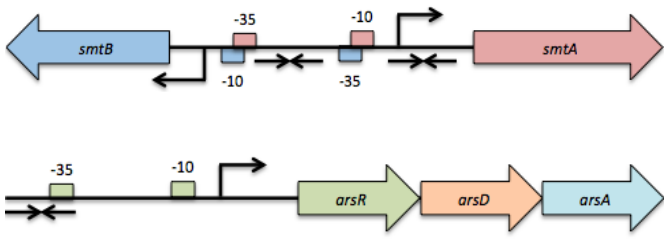
Data from the above binding sites were collated and compared to define typical 12-2-12 DNA sequences to which SmtB-like (bind to different strands) and ArsR-like (bind to same strand) members of the ArsR family bind, as shown in Figure 1.6a and 1.6b [28]. Figure 1.6c shows the location of the DNA binding sites with respect to the genes encoding for both the ArsR family protein and the associated resistance system for ArsR and SmtB. The 12-2-12 region is either directly overlapping the promoter site or is between it and the resistance gene, blocking expression by RNA polymerase. It is also worth noting that many regulators are self-regulating,

5 - aATAxxTGAacaxxtatTCAxaTxtt - 3

(a)

5 - TAxAxTCAAAtaxxtaTTTGaxTxTA - 3

(b)



(c)

Figure 1.6: Conserved 12-2-12 DNA sequences to which (a) SmtB- and (b) ArsR-like members of the ArsR family bind, classified with regard to whether the protein binds to both strands of DNA or just one. Capitalised bases are more conserved than lower case bases [28]. (c) Locations of DNA binding sites for SmtB and ArsR, indicated by →←, in relation to the promoters for the genes whose transcription they regulate. Based on [28].

with the gene encoding for both it and the protein it regulates being divergent within the gene cluster, such that when the ArsR family protein binds to DNA it inhibits the promoter regions for both genes. This is thought to be the case for approximately two thirds of ArsR family proteins [71].

1.2.4 Ligand Interactions

It is well documented that members of the ArsR family can sense a range of metal ions, with some ArsR proteins capable of binding several different metal ions, whilst others can bind only one [51]. The range of metals that have been demonstrated to bind to ArsR proteins are shown in Table 1.3, along with examples of which family members bind each metal. No obvious groupings can be formed of proteins which bind similar metals, or metals that are bound by a similar group of proteins. However, in general, it is thought that family members with an α3-based metal binding site are responsible for sensing heavy metal pollutants in the cellular environment, whereas those with an α5-based site tend to sense biologically required metal ions [28]. As such, α5 sites typically have less sensitivity to their metal ligands, as a small amount of the ion is not detrimental (and is often vital) to the cell's survival, whereas α3 sites have a much stronger affinity so that the system can detect even minute amounts of pollutant heavy metal ions [51].

An ensemble of 554 of the previously mentioned 1024 ArsR sequences was created using CLANS [79]; excluding the most distant relatives [57], a tree diagram of which can be seen in Figure 1.7. It is apparent that there is no distinct grouping of the different sensing sites and

Metal	**Sensor Example**
Antimony	ArsR [46]
Arsenic	ArsR [46]
Bismuth	ArsR [7]
Cadmium	AztR [53], BxmR [72], CadC [47], CmtR [73]
Cobalt	CzrA [74], KmtR [57], NmtR [75]
Copper	BxmR [72]
Lead	AztR [76], CadC [77], CmtR [73]
Nickel	KmtR [57], NmtR [75]
Silver	BxmR [72]
Zinc	AztR [76], BxmR [72], CadC [77], CzrA [74], SmtB [52], ZiaR [78]

Table 1.3: List of metal ions known to be sensed by different ArsR family member proteins, along with examples to which each binds. NB: CzrA is also known as ZntR in the literature [28].

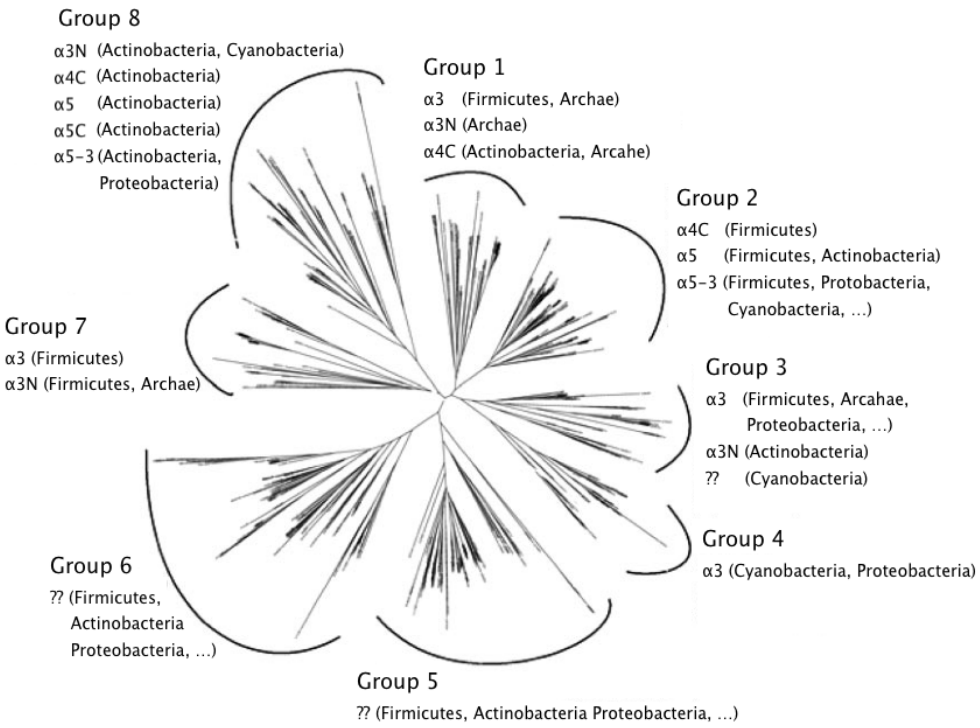


Figure 1.7: Tree diagram of the 554 closest ArsR family sequences from the Pfam database [56]. "??" indicates categorisation of ArsR proteins that do not contain any of the patterns indicated in Table 1.2. Main bacterial phyla are shown in parenthesis, with elipses indicating sequences that are present in other phyla. Modified from [57].

that proteins from the same phylum that have the same sensing site can appear in different groups. It is, therefore, difficult to conclude that all family members with, for example, high similarity to ArsR will bind arsenite and antimonite. There is not a one to one correlation between metal binding sites and the sequence groups that are assigned here.

The issue is complicated further by the identification of a sub-group of the ArsR family that regulate genes that are not involved in the regulation of metal ions within the cell. Two examples of this sub group are EcaR, from *Erwinia caratovora*, and HlyU, thought to regulate expression of haemolysin (HlyA) in *Vibrio cholerae*. These have both been shown to specifically bind to DNA in their own operator regions, but neither of which could be dissociated from the DNA by the addition of any of the metal ions ArsR proteins are known to bind to [57]. As each of these proteins contains a cysteine in both α2 and α4, it is thought that this could indicate another binding motif additional to those shown in Table 1.2, the α2α4 motif, predicting non-metal sensors [51].

1.2.5 Structural Properties

It has been alluded to several times thus far, especially in Section 1.2.1, that ArsR family proteins have a strong helical character, beyond the two helices needed for the HTH DNA binding domain. This was first demonstrated when the crystal structure of SmtB was solved, identifying five α-helices and a double stranded β-sheet, arranged as an α1-α2-α3-αR-β1-β2-α5 fold, where αR designates the recognition helix of the HTH domain, along with α3 [80]. The full dimeric structure is shown in Figure 1.8, visualising data from PDB entry 1SMT [80] using Pymol [42].

It can be seen from Figure 1.8 that the dimeric interface of SmtB forms via interactions between the pairs of α1 and α5 helices, resulting in a configuration with both recognition helices presented on the same side of the dimer for DNA binding. The α3-turn-αR sub-structure can also be directly compared to the typical HTH structure shown in Figure 1.2 (comprising α2 and α3 in that instance). From this structure, it can also be seen how the different metal binding sites identified in Table 1.2 lead to different mechanisms of action. For instance, it is easy to see that a metal ion binding to the α3 helix would directly interfere with DNA binding [81], whereas binding to α5 would alter the dynamics of the dimer interface, resulting in a shift in conformation with a direct impact on DNA affinity [80]. From this and later structures, the ArsR family have sometimes been assigned as winged HTH in character (a subgroup of HTH transcription factors [82]) due to the positioning of the β-sheet with respect to the DNA binding domain. However, it is often the case that winged HTH proteins also interact with

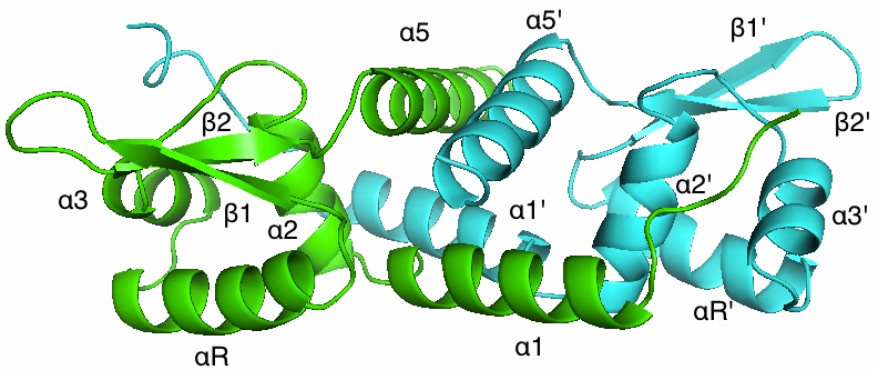


Figure 1.8: Crystal structure of SmtB visualised in Pymol using data from [80]. α-helices and β-sheets are numbered starting at the N terminal, with green and cyan representing different monomers.

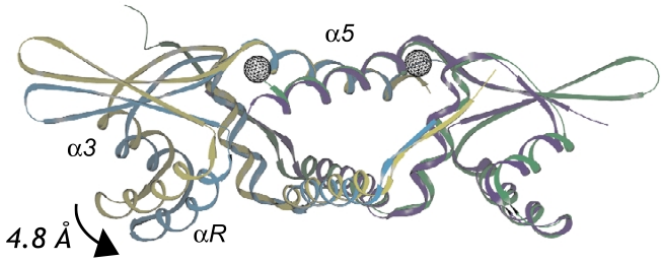


Figure 1.9: Overlay of crystal structure of Zn_2 SmtB dimer (purple and blue) with apo-SmtB dimer (green and gold). Right hand monomers are aligned to each other, resulting in an observable shift in relative position of the recognition helix of the second Zn bound monomer by 4.8 Å. Zinc binding sites also clearly highlighted in α5-helices. Taken from [84].

minor grooves of DNA through their β-sheets, whereas this is only the case for a few ArsR proteins [83], hence the ArsR family can perhaps be thought to exist between the HTH and winged HTH groups.

Figure 1.9 shows an overlay of the crystal structure of zinc bound SmtB with the apo-form, clearly demonstrating the structural shift leading to a reduction in DNA affinity [84]. It can be seen that upon binding two zinc ions to the α5 helices the relative position between αR helices is altered by 4.8 Å, enough to cause misalignment between the two recognition helices and the

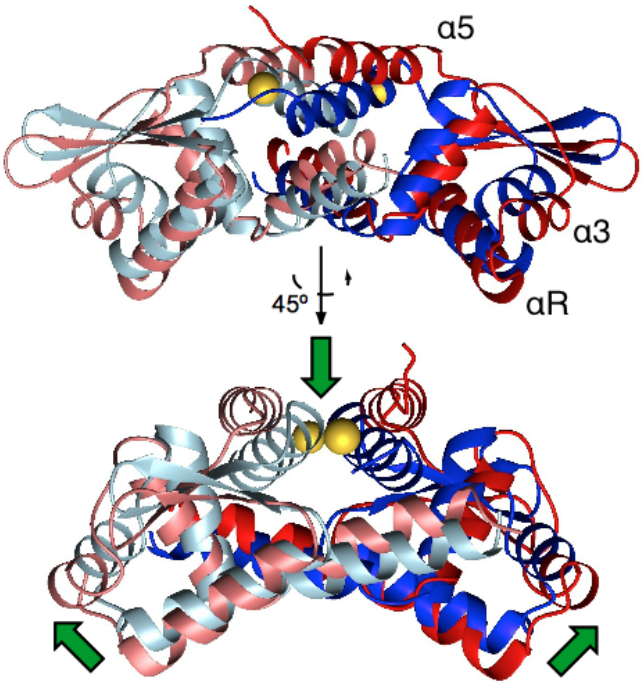


Figure 1.10: Overlay of NMR structure of DNA-bound CzrA (red) and crystal structure of Zn(II) bound CzrA (blue), viewed from two different perspectives. Bold and pale colouring indicates the individual monomers in the dimeric complexes, and Zn(II) is illustrated as gold spheres. Green arrows illustrate the conformational change in the two structures. Taken from [81].

major grooves of the target DNA.

The functional shift in conformation of α5-binding ArsR proteins can be seen more explicitly in Figure 1.10, which shows structural data for CzrA obtained by solution Nuclear Magnetic Resonance (NMR) when bound to DNA as well as the crystal structure of Zn(II) CzrA [81]. In this instance, the binding of Zn^{2+} to the two α5 helices draws these helices closer together and also deeper into the centre of the complex, as shown by the top green arrow. This then leads to further conformational change, resulting in a shift in the relative position of the two HTH domains to a more open conformation, similar to that seen in SmtB in Figure 1.9. Again, this shift in conformation causes misalignment between the two HTH motifs and the major grooves of the DNA target, reducing binding affinity and causing dissociation with DNA upon Zn(II) binding. In this case, the global backbone root mean squared deviation (RMSD) between the two structures is 2.8 Å [81]. However, it is noted that a quantitative comparison at this level may be invalid due to the possible inherent deviation between the DNA and Zn(II) bound structures as they were determined by different techniques. Further work has been carried out to provide a more accurate comparison via molecular dynamics simulations [85].

While no direct structural comparisons of DNA-/metal-bound α3 sites have been reported in the literature,[1] an NMR structure showing cadmium bound to the α4C site in CmtR has been published [86]. It has been demonstrated that this metal site does indeed include cysteine residues in the recognition helix of CmtR, as well as Cys102 in the C terminal region. This locks the recognition helix and C terminus into close proximity, reducing the mobility of the protein such that it cannot adopt a conformation amenable to DNA binding. This is displayed in Figure 1.11, demonstrating a similar yet distinct method of action when compared to binding of the α5 site in other family members.

Another insight into ArsR family structure can be gained from the crystal structures of two HlyU orthologues: HlyU-Vv from *Vibrio vulnificus* CMCP6 [87] and HlyU-Vc from *Vibrio cholerae* N16961 [83]. It was found that the dimeric interface in HlyU-Vv differed slightly from that of SmtB by the addition of an extra polar interaction between the α5-helices, as well as deviations in length and orientation [87]. It was also noted that the distance between α1 and

[1] It should be noted that apo- and metal-bound CadC structures have been reported [77], but these are with Zn(II) bound to the α5 site, rather than either Cd(II) or Pb(II) bound to the α3 site.

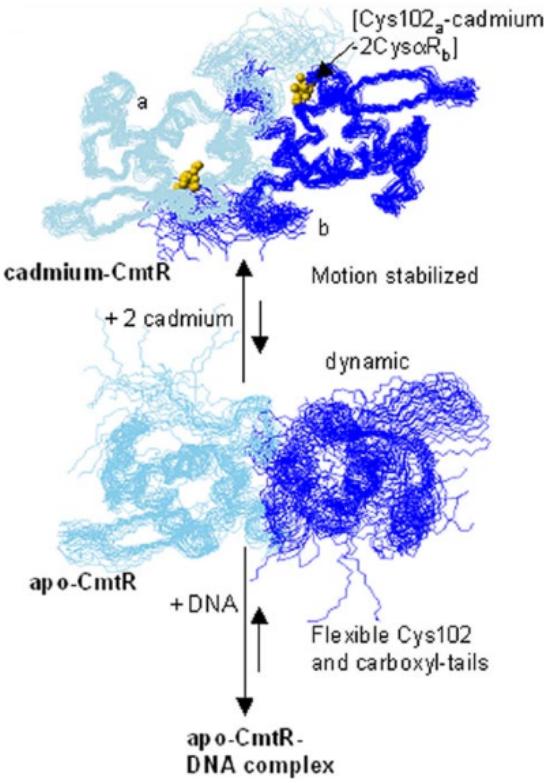


Figure 1.11: Illustration of reduced flexibility in CmtR upon binding of cadmium ions to α4C binding site. Apo-CmtR structure set results from molecular dynamics simulations using the CYANA algorithm based on NOE measurements of Cd-CmtR, whereas the Cd-CmtR structures themselves are the 30 lowest energy backbone conformations experimentally determined by NMR. Taken from [86].

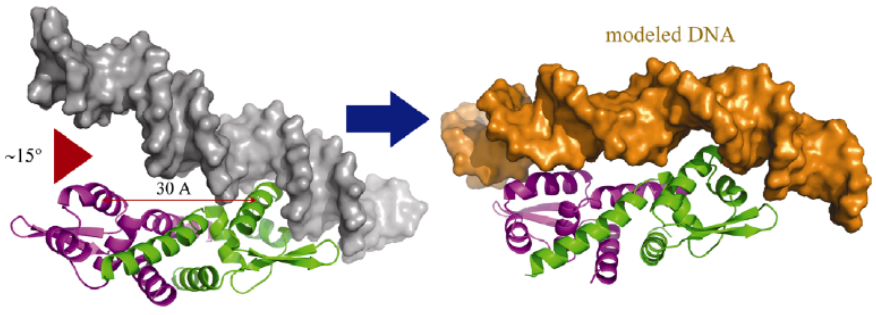


Figure 1.12: Model of DNA binding by HlyU-Vv, indicating that a 15° bend of the DNA is required to align with the two recognition helices. The two monomers of HlyU-Vv are coloured green and purple. Taken from [87].

αR in HlyU-Vv was larger than in SmtB, with the distance between αR helices reduced to 30 Å compared to 34 Å in SmtB [83]. This indicated that DNA would have to bend by approximately 15° for binding by HlyU-Vv, as shown in Figure 1.12. With this being the case, it is easy to see how, if the predicted α2α4 cysteine ligand binding model is correct, any further interactions and stresses around the recognition helix α4 would lead to DNA dissociation.

HlyU does not contain any typical ArsR metal binding sites, instead having cysteines present in α2 and αR/4 as previously mentioned, and has been shown to be non-responsive to any metal ligand known to bind to ArsR family members [57]. Work carried out on HlyU-Vc has identified that Cys38 in helix α2 can be modified to sulfenic acid, hinting at the possibility that this protein operates as a redox switch [83], similar to the MerR family protein SoxR [58] and the ArsR family protein BigR (see below). It is interesting to note that the αR-helices in HlyU-Vc are even closer together than those in HlyU-Vv, only 27 Å apart, meaning a DNA bend of 68° is required for major groove alignment. This is possible in HlyU-Vc as it brings the two β-sheets into contact with the minor grooves, making HlyU-Vc one of the ArsR family proteins that can

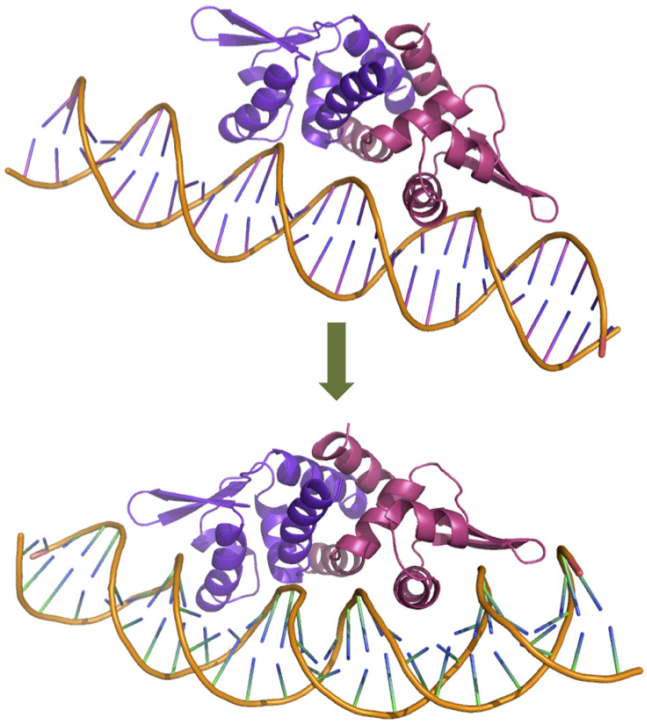
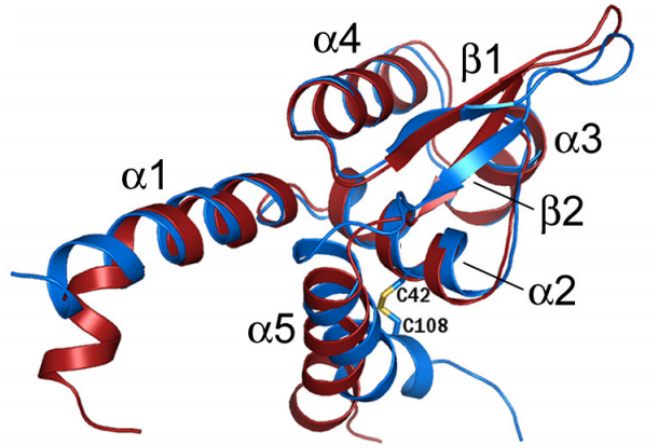
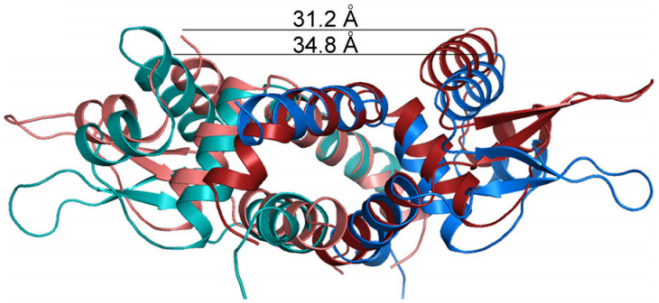


Figure 1.13: Model of crystal structure of HlyU-Vc interacting with DNA, with purple and pink representing the individual monomers. It can be seen that the β-sheets of the dimer align with the minor grooves of the DNA, stabilising the bond despite the 68° bend required to align the major grooves with the recognition helices. Taken from [83].



(a)



(b)

Figure 1.14: Crystal structure of BigR. (a) Monomer, with disulphide bridge highlighted as yellow sticks. Red shows the reduced form while blue shows the oxidised form. (b) Dimer in both reduced (red) and oxidised (blue) form, with the difference in distance between recognition helices highlighted. As with other ArsR family proteins, this difference is on the order of 4 Å. In this alignment, the α1 helices of both forms have been superimposed. Modified from [88].

truly be classified as winged HTH, as seen in Figure 1.13.

Finally, BigR is another example of a winged HTH ArsR family protein which has in this case been shown conclusively to act as a redox switch [88]. It has been demonstrated that the formation of a disulphide bridge between C42 and C108 locks the protein in a conformation which is not amenable to interactions with DNA, as demonstrated in the solved crystal structures shown in Figure 1.14. Hence it can be seen that when the cell experiences hypoxia, the BigR dimer will dissociate from DNA, allowing transcription of the regulated genes which, in this case, encode for a hydrogen sulphide detoxification mechanism.

All of the structures presented here indicate that while there are certain consistent elements to the structures of ArsR family proteins, there are also characteristics specific to each of the family members whose structure has been solved. This adds weight to the previous assignment of at least 8 sub groups being present within the family, as predicted by database analysis

in [50].

1.3 A Novel ArsR-like Protein Regulating Antibiotic Export?

1.3.1 The Methylenomycin Gene Cluster

Methylenomycin (Mm) A was first characterised in 1974 as an antibiotic produced by *Streptomyces violaceoruber* SANK 95570, active against both Gram-positive and Gram-negative species [89], although their mechanism of action is still unknown. Further work into the biosynthetic pathway also led to the identification of Methylenomycin C [90], the structure of which, along with MmA, is shown in Figure 1.15. It was found that the Methylenomycins were produced by *Streptomyces coelicolor* A3(2) by genes on the SCP1 plasmid, making them the first antibiotic compound whose synthesis and resistance was entirely plasmid determined [91, 92].

These species were found to be resistant to the Mm they produced via the Mmr protein [93, 94], later found to operate as a membrane integral efflux pump through comparison with the known efflux protein Tcr, which provides resistance to tetracycline in *E. coli* [95]. The *mmr* gene was found to be embedded within a cluster of biosynthetic genes, leading to the possibility that the genes encoding the resistance mechanism may only be transcribed during biosynthesis of MmA [96]. Plasmid SCP1 has been sequenced in its entirety [97], allowing the Methylenomycin gene cluster to be fully annotated, a schematic of which is shown in Figure 1.16a. It has been noted that there is an imperfect inverted repeat within the intergenic region between the *mmr* gene and its divergent neighbour, *mmyJ*, suspected of being a possible transcription factor binding site [95]. Further investigation indicated that this region, shown in Figure 1.16b, was indeed protected from RNA polymerase. As it captures the promoter regions for both *mmyJ* and *mmr* genes, it was suspected that the protecting protein was a transcriptional repressor,

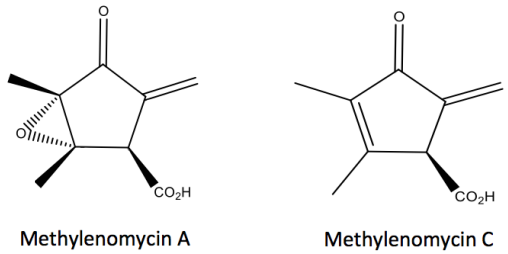
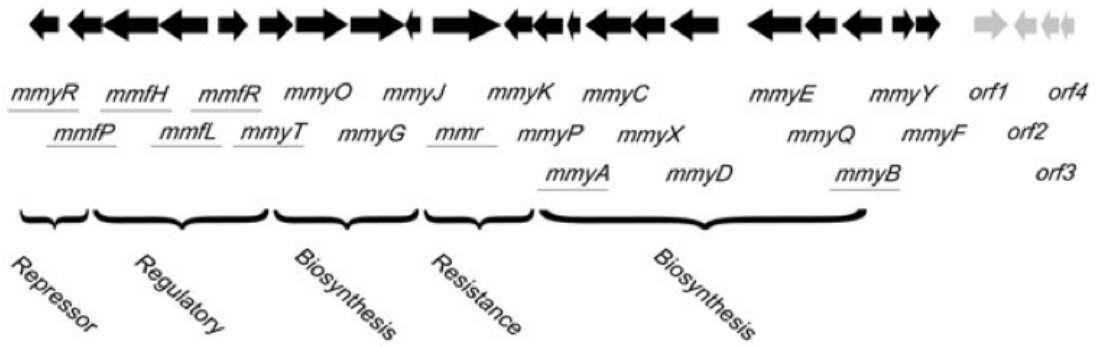
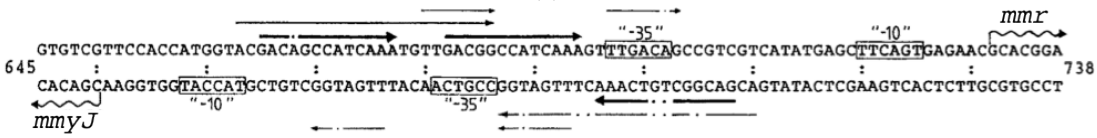


Figure 1.15: Structures of Methylenomycin A [89] and C [90].



(a)



(b)

Figure 1.16: (a) Methylenomycin gene cluster from SCP1 plasmid. This cluster is pictured in accordance to orientation previously described in the literature [96], even though the sequenced genome was orientated the other way around [97]. Taken from [99]. (b) Promoter regions of *mmyJ* and *mmr*. Straight arrows indicate repeated sequences, with dots highlighting mismatches. Colons mark every tenth base pair and wavy arrows indicate mRNA transcription origins. Sequence originally reported in [95]. Modified from [100].

encoded by the *mmyJ* gene. It is also known that *mmr* transcription begins upon production of MmA [98], and so we propose here that MmyJ is a transcriptional regulator, sensitive to MmA.

1.3.2 MmyJ: A Novel ArsR Family Protein

MmyJ is encoded by a relatively short gene (336 bp) divergent to *mmr* and separated by just 218 bps. This position, along with the previously discussed protected region of the intergenic DNA containing the promoter sequences for both *mmyJ* and *mmr*, indicates that MmyJ could be a self-regulating transcription factor, which also regulates expression of the Mmr efflux pump.

MmyJ is a 111 amino acid/12.1 kDa protein, encoded for by the *mmyJ* gene as displayed in Figure 1.17. An analysis of the resulting amino acid sequence by NCBI's pBLAST (protein Basic Local Alignment Search Tool) [101] returned an ArsR-like classification (see Chapter 2 for more information). Through observation of the amino acid sequence shown in Figure 1.17, however, it can be seen that this classification is slightly surprising, as there is only a single cysteine present in the protein at residue 49, and, as previously discussed, cysteine residues typically form the main ligand site in ArsR proteins. Interestingly, based on predicted protein structure (Chapter 2), this cysteine lies on the edge of the α3 helix, deviating from the observed

```
  1  GTG GCG GCA CGG ATC ACG ACA GAG CGC ATC ACC GAC CAT CCG GAC GCT GAC
       M   A   A   R   I   T   T   E   R   I   T   D   H   P   D   A   D

 52  GCC ATC ACC CTC CAG GGC GTC CTG GAC GCG CTG GTC GAT CCG GTG CGC CGC
       A   I   T   L   Q   G   V   L   D   A   L   V   D   P   V   R   R

103  AGC ATC GTC CGG CAG CTG GCT AAG GCA CCC GAG GAC ATC GCC TGC GGC ACC
       S   I   V   R   Q   L   A   K   A   P   E   D   I   A   C   G   T

154  TTC GAC ATC ACC GTC TCC CGC TCG ACC GGC ACT CAC CAC TTC AAG GTG TTG
       F   D   I   T   V   S   R   S   T   G   T   H   H   F   K   V   L

205  CGC CAG GCC GGG ATC ATC AGG CAG TAC TAC ATC GGC ACC TCG AAG ATG AAC
       R   Q   A   G   I   I   R   Q   Y   Y   I   G   T   S   K   M   N

256  ACG CTT CGC ACC GAT GAT CTC GAT CAG GCC TTC CCC GGC CTG CTC ACC GCG
       T   L   R   T   D   D   L   D   Q   A   F   P   G   L   L   T   A

307  ATC GTC GAC GCC GCG GCC AGG GAG AGC TGA
       I   V   D   A   A   A   R   E   S   -
```

Figure 1.17: Genetic sequence and corresponding amino acid sequence for MmyJ [97].

pattern of suspected non-metal binding ArsR proteins containing the α2α4 motif. However, as this system is regulating Mm production, and is not proposed to involve metal sensing, it is possible that cysteine residues are not needed if MmA or one of its biosynthetic intermediates is sensed by MmyJ. If this is the case it would be the first reported instance of an ArsR family protein sensing a non-metallic ligand.

1.4 Project Aims & Objectives

The main aim of this project is to investigate MmyJ and determine its structure and function, and hence conclude whether it is truly a novel ArsR-like transcriptional repressor and is the first instance of an ArsR family protein known to bind an organic molecule, rather than a metallic ligand. An in depth bioinformatics analysis will be carried out, alongside experimental work with expressed protein and amplified fragments of the *mmr-mmyJ* intergenic region to determine binding characteristics. Work will also be carried out into identifying the binding ligand and/or conditions that cause dissociation of MmyJ from DNA if it is proven to be a transcription factor. Also, attempts will be made to confirm that the identified imperfect inverted repeat surrounding the *mmyJ/mmr* promoter regions is in fact the binding site of MmyJ, confirming its role as a transcriptional repressor of itself as well as *mmr*. Once the DNA

target region and ligand have been identified, surface plasmon resonance (SPR) will be used to determine the binding kinetics of the interactions.

As well as the above functional investigation, work will be carried out to determine MmyJ's structure by solution and solid state NMR, as well as X-Ray crystallography if conditions can be determined that lead to crystallisation of the protein. From this, work will then be carried out, if possible, to observe the structure of MmyJ when bound to both DNA and its ligand, with the hope that direct comparisons between these structures and that of apo-MmyJ will lead to insights into the specific residues needed for each interaction. Data acquired through NMR will also be used to complement the analysis by SPR, investigating the link between residue mobility and binding kinetics at the interaction sites.

2 Bioinformatic Analysis of MmyJ

Before characterising MmyJ by experimental means, bioinformatic studies were carried out to investigate MmyJ and determine its similarities and differences to other ArsR proteins.

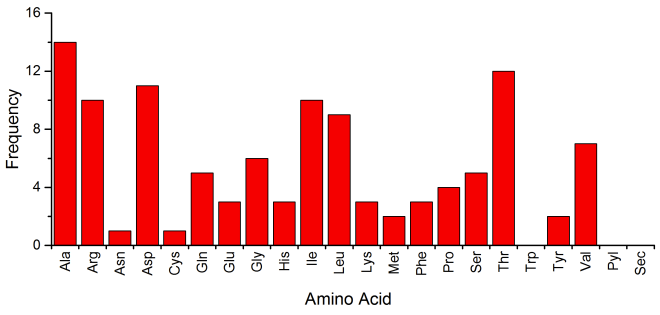
2.1 MmyJ - Basic Properties

The amino acid sequence for MmyJ is shown in Figure 2.1. This was first determined by forming a cosmid library based on the linear plasmid SCP1 from *S. coelicolor* A3(2), sequencing it and translating the coding sequences (CDS) from these cosmids. Specifically, cosmid C73 contains the entire methylenomycin pathway [99]. As can be seen, the resulting protein corresponding to the *mmyJ* gene is 111 amino acids long, with a corresponding mass of 12135.8 Da. The amino acid composition can be seen in Figure 2.1b. This sequence was then entered into the ProtParam [102] web tool in order to calculate basic physical properties of the protein, relevant for later work. One such property calculated was the extinction coefficient, found to be 2980±300 $M^{-1}cm^{-1}$, corresponding to an absorbance at 280 nm of 0.246±0.02 when at a concentration of 1 g/L.[2] This allowed the concentration of MmyJ to be determined when produced in the laboratory (although a modification to 0.281±0.03 was required for recombinant MmyJ with histidine tag, as described in Section 3). Also, the computations indicated that

[2]Calculated uncertainty of 10% in approximation of extinction coefficient is due to lack of Trp residues [102].

```
        10         20         30         40         50         60
MAARITTERI TDHPDADAIT LOGVLDALVD PVRRSIVROL AKAPEDIACG TFDITVSRST

        70         80         90        100        110
GTHHFKVLRQ AGIIRQYYIG TSKMNTLRTD DLDQAFPGLL TAIVDAAARE S
```

(a)



(b)

Figure 2.1: (a) Amino acid sequence and (b) amino acid composition of MmyJ, determined by coding sequence (CDS) analysis.

MmyJ may be unstable, with a calculated instability index of 45.39, and so measures were taken when expressed to improve stability, such as ensuring the protein was kept in solution containing 10% glycerol (see Chapter 3 for more information regarding protein handling after expression).

2.2 Motif Recognition & Alignment

2.2.1 BLAST Analysis & Classification

Initial classification of MmyJ as an ArsR protein was carried out by entering the protein sequence shown in Figure 2.1 into BLAST (Basic Local Alignment Search Tool) [101]. Figure 2.2 shows the conserved domains detected using the BLASTp algorithm[3] to compare the amino acid sequence of MmyJ to the NCBI nr database[4] [103, 104, 105, 23]. It can immediately be seen that there is a specific hit identifying residues 25-83 as an ArsR helix-turn-helix (HTH) motif. The motif was assigned with an expected value of 1.10×10^{-6}, giving the result moderately high significance (expected value indicates the likelihood that the motif is recognised purely by chance rather than there being a true assignment). However, the lowest expected value belongs to the assignment of the general ArsR identification of residues 8-103, with a value of 8.43×10^{-10}, and so even if the exact assignment of the HTH_ARSR motif is incorrect, it is highly likely that MmyJ is still an ArsR family protein. The other non-specific hits, the second possible HTH_ARSR and HTH_5 domains, have expected values of 1.19×10^{-9} and 9.9×10^{-7} respectively, offering alternative assignments in the case of the specific hit proving to be inaccurate. The closeness in expectation values of these possibilities implies that while it is likely that there is a HTH motif within the area identified as being ArsR-like, its precise location perhaps cannot be determined by this method.

[3] BLASTp is a form of BLAST optimised for comparing protein sequences.
[4] The nr, or non-redundant, database contains sequences from GenBank translations as well as Refseq, PDB, SwissProt, PIR and PRF databases with duplicate sequences removed.

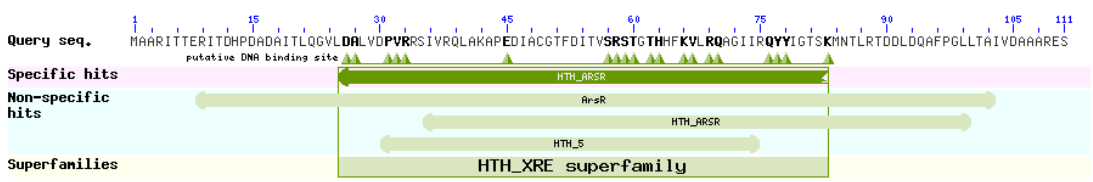
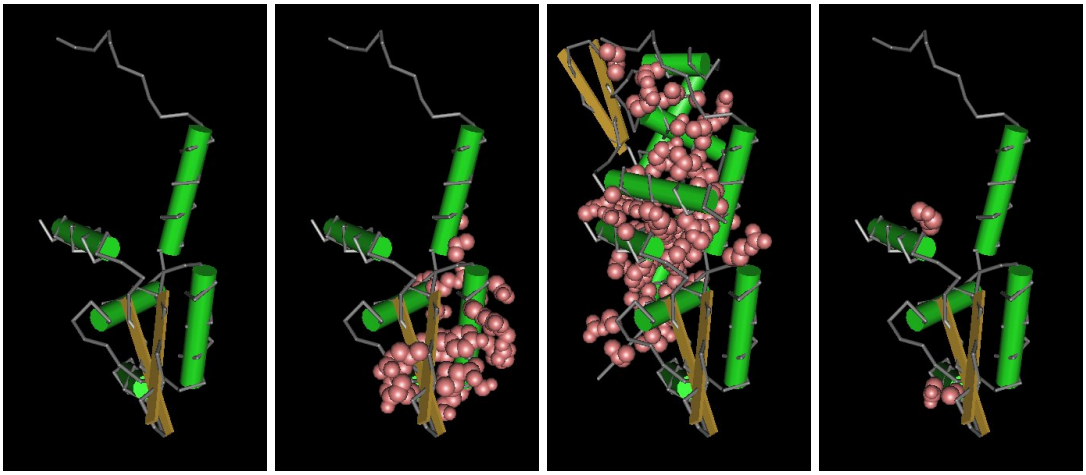


Figure 2.2: Graphical summary of conserved domains within MmyJ, identified using NCBI BLASTp.



(a) (b) (c) (d)

Figure 2.3: Crystal structure of SmtB viewed in Cn3D. The isolated monomer is shown in (a), with pink balls in (b), (c) and (d) representing putative DNA binding sites, dimeric interfaces and Zn^{2+} binding sites respectively [80].

However, it is still worth looking closer at the specific hit indicated in Figure 2.2, as part of this assignment includes the identification of 20 putative DNA binding sites out of the 24 that are thought to be conserved as part of the motif, specifically those identified in SmtB [80, 23]. Figure 2.3 displays the position of these putative DNA binding sites on the structure of SmtB, along with dimer interface region and Zn^{2+} binding sites. The predicted structure of MmyJ is compared to SmtB in Section 2.3, and so it is discussed there whether the highlighted DNA, dimer and Zn^{2+} binding sites could share common features to that of MmyJ.

Another aspect of the BLAST output is statistical analysis of the degree of overlap, identity and similarity between a queried sequence and the closest protein sequences in the database, with the default being the top 100 matched; the top 10 of which are shown in Table 2.1. While some caution should be taken when using this analysis (for example if the 5 residues on the N-terminal of the query sequence match the 5 residues on the C-terminal on a sequence in the database, the analysis will return 100% identity but only for a small degree of overlap), it is possible to extract useful information. In this instance, it was discovered that of the 100 closest protein sequences in the database, 80 were classified as ArsR proteins, with all others being hypothetical proteins with no assigned function at the time of analysis, further cementing the conclusion that MmyJ is indeed an ArsR protein. Of these 80 sequence alignments, all but one had 50% or higher identity over 90 or more amino acids of the MmyJ sequence. The average identity over all 80 sequences was 58.87% over 90 or more amino acids, with the average

Description and Species	% ID	% Cover	Accension #
ArsR family transcriptional regulator *Streptomyces* NRRL S-31	92	100	WP_030742020.1
hypothetical protein *Streptomyces* 351MFTsu5.1	75	99	WP_030742020.1
ArsR family transcriptional regulator *Amycolatopsis* (Multispecies)	72	94	WP_020419963.1
ArsR family transcriptional regulator *Streptomyces niger*	72	90	WP_052868623.1
ArsR family transcriptional regulator *Streptomyces* (Multispecies)	70	100	WP_030286058.1
hypothetical protein SBI_00765 *Streptomyces bingchenggensis* BCW-1	70	83	ADI03886.1
ArsR family transcriptional regulator *Streptomyces mirabilis*	70	95	WP_037721374.1
ArsR family transcriptional regulator *Kitasatospora mediocidica*	69	96	WP_035791758.1
ArsR family transcriptional regulator *Gordonia soli*	68	88	WP_007625248.1
hypothetical protein *Streptomyces nanchangensis*	68	90	ADC45569.1

Table 2.1: Top ten hits when comparing the sequence of MmyJ to the nr database using BLASTp, listed by percentage identity (ID). All can be seen to have high identity and cover. The top two hits are thought to be orthologues of MmyJ (see Section 2.4).

similarity (allowing for substitution of amino acids for others with similar properties) found to be 76.08% over the same range. This indicates a very high degree of homology with a significant number of ArsR proteins. It is also worth noting that all of the expected values for the alignments included in the above analysis were of the order 10^{-31} or lower and so it can confidently be stated that these represent true alignments statistically.

2.2.2 Prosite Comparison with other ArsR Family Proteins

In order to further compare MmyJ to similar proteins, the sequences of SmtB, ArsR and 7 other ArsR family proteins were selected from those mentioned in [51], details of which can be seen in Table 2.2. These proteins were selected as they are produced by a range of species, with 4 produced by Gram-negative and 5 by Gram-positive organisms, as well as sensing a range of metals, binding between 0 and 4 of the 10 metals known to bind to ArsR proteins. In this way it was hoped to attempt a broad comparison across the ArsR family.

Amino acid sequences for the proteins shown in Table 2.2 were obtained from the GenBank database [110] hosted by NCBI. These, along with the sequence for MmyJ, were then entered into ExPASy's ScanProsite tool [107], which compares user defined sequences with a database

Name	Organism	GenBank ID	Metals Sensed [51]
SmtB	*Synechococcus elongatus* PCC 7942 (-)	CAA45872.1	Zn
ArsR	*Acinetobacter baumannii* (-)	AFC76435.1	As, Sb
ZiaR	*Bacillus thuringiensis* Bt407 (+)	AFV16327.1	Zn
BxmR	*Oscillatoria brevis* (-)	BAD11074.1	Cu, Ag, Cd, Zn
CadC	*Staphylococcus aureus* (+)	AEH58925.1	Cd, Pb, Zn
CzrA	*Bacillus amyloliquefaciens* (+) TA208	YP_005541188.1	Zn, Co
HlyU	*Vibrio cholerae* (-)	CAA47336.1	None
KmtR	*Mycobacterium tuberculosis* H37Rv (+)	NP_215342.1	Ni, Co
CmtR	*Mycobacterium tuberculosis* H37Rv (+)	CCP44766.1	Cd, Pb

Table 2.2: Details of proteins included in a test set of ArsR family proteins for comparison with MmyJ. (+) and (-) at the end of each species denotes its response to the Gram staining test, as listed in [106]. Note: the organism listed corresponds to the orthologue of each protein as listed in Figure 2.4.

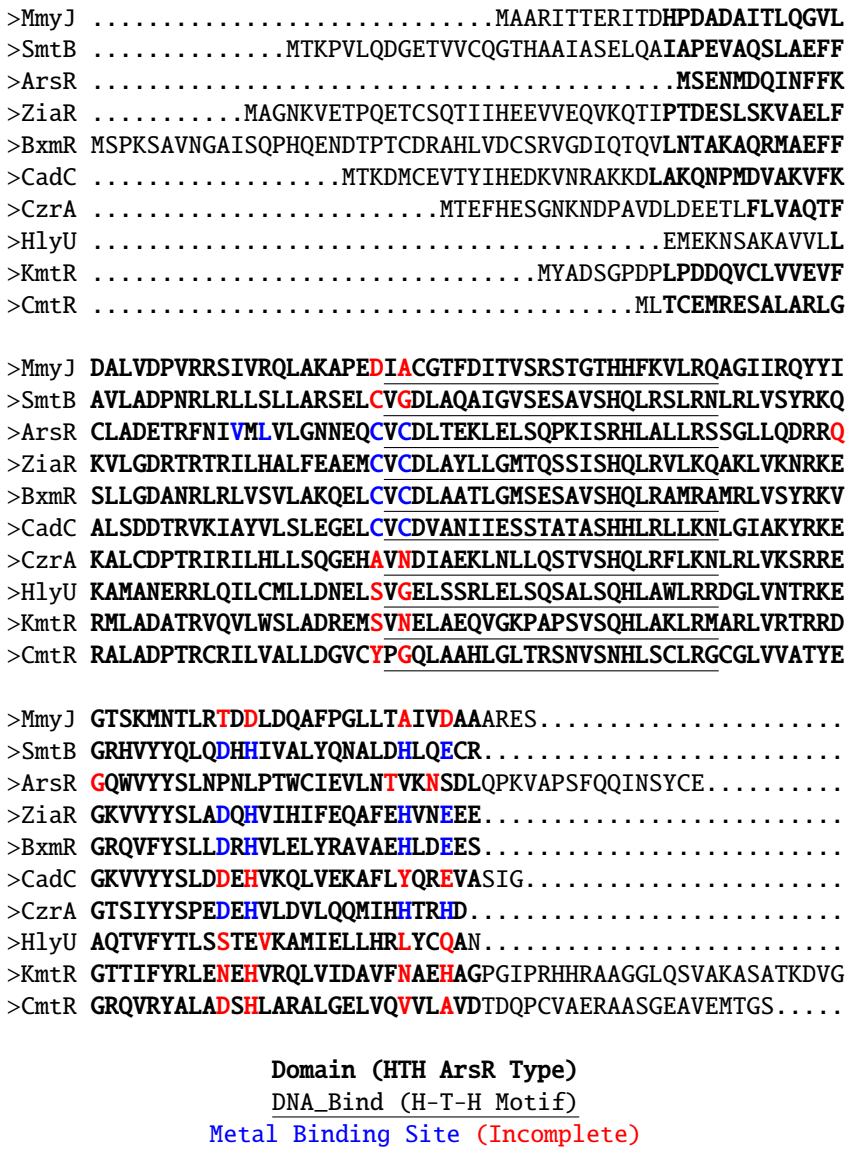


Figure 2.4: Predicted domains, motifs and metal sites in MmyJ, SmtB, ArsR and 7 other selected ArsR family protein sequences. Red residues indicate predicted metal binding sites thought by the algorithm to be incomplete. Analysis was carried out using ScanProsite [107] and the PROSITE database [108, 109].

[108] according to PROSITE methodology [109] in order to identify motifs and functional sites. The result of this analysis can be seen in Figure 2.4, and clearly there is a conserved motif between all 10 protein sequences: namely that identified as DNA_Bind (H-T-H Motif). Whilst all test proteins are also identified as containing the generic HTH ArsR Type Domain, the length of this varies between proteins, as does the position of the DNA binding motif within this generic domain. As such, it is concluded that the DNA binding motif is the key part of the ArsR domain as defined by PROSITE, and hence this is used as further evidence of MmyJ being identified as an ArsR type protein. It is interesting to note that all metal sites predicted to be part of an ArsR motif by PROSITE are absent in the amino acid sequence of MmyJ, possibly indicating that it is not a metal binding protein. However, the same can be said for HlyU, KmtR and CmtR and, while HlyU has been shown not to bind to metal ions, KmtR and CmtR are both known to be metal binders, hence no adequate conclusion can be drawn from their predicted absence from MmyJ.

By comparing this PROSITE analysis to the previous one done with BLAST, it can be seen that the HTH_ArsR domain, identified by BLAST, corresponds to just a section of the HTH ArsR Type domain, identified by PROSITE, although it also contains the entire DNA_Bind (H-T-H Motif) sequence, indicating a high level of confidence in the presence of an ArsR type HTH domain within MmyJ. However, the two analyses are slightly contradictory in their prediction of DNA binding sites - only 10 of the 20 identified by BLAST in Figure 2.2 coincide with the HTH motif as identified by PROSITE, demonstrating how these analyses cannot be simply taken at face value.

2.2.3 MEME Analysis

The amino acid sequences identified as comprising an ArsR family DNA binding HTH motif by PROSITE were extraced and analysed further using the MEME online tool [111]. The result of this analysis can be seen in Figure 2.5. While there are no instances of a single amino acid being conserved across all 10 proteins, there are at least several points where amino acid type is conserved, for example residues 8, 19 and 22 being conserved as hydrophobic.

It is apparent from Figure 2.5 that, while MmyJ has the highest p-value (equivalent to the expected value previously explained for BLAST) and therefore can be thought of as the least

Figure 2.5: Graphical output showing conservation of amino acids (both individually and grouped by types) from the MEME online tool [111]. Blue indicates hydrophobic residues, green indicates polar, non-charged, non-aliphatic residues, magenta indicates acidic residues and red indicates positively charged residues. Pink, orange, yellow and turquoise indicate ungrouped residues. Grey indicates that residues are not identified as being part of the motif. Proteins are ranked in order of ascending p-value.

likely protein sequence out of the 10 submitted to actually contain this motif, there is still a high degree of alignment between it and the other proteins tested. It is logical that there is some variation within this motif as each protein is likely to bind to a different sequence of DNA and, as such, it is still felt that MmyJ contains this motif with a high significance. Hence, it can be concluded that MmyJ is an ArsR family protein and does appear to contain a typical HTH DNA binding domain.

2.3 Homology Modelling

Having established with confidence that MmyJ is an ArsR-type protein, homology modelling was carried out to approximate its 3D structure.

2.3.1 Monomer

The amino acid sequence for MmyJ was entered into the Phyre2 web portal [112]. This algorithm finds homologues using PSI-BLAST[5] [113], before constructing models based on structural information found for these homologues and iterating these models until the resulting structure is of sufficiently high quality. Figure 2.6 shows the best predicted 3D structure of MmyJ from

[5]PSI-BLAST differs from the previously used BLASTp algorithm in that it uses position specific scoring matrices, unique to each sequence comparison, rather than the pre-defined scoring matrices traditionally used in BLAST. This increases sensitivity of the search but also increases computation time.

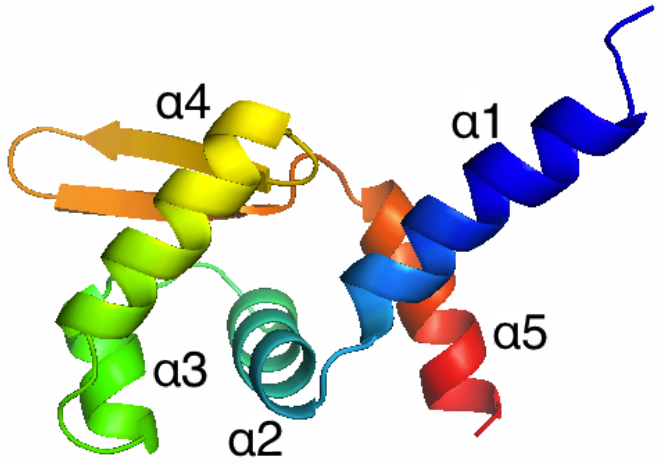


Figure 2.6: Phyre2 homology model of MmyJ based on template c2lkpB, the solution structure of NmtR. This structure models 93 residues (84% of MmyJ sequence) with a confidence of 99.9% or higher. Chain is coloured such that N terminus is blue and C terminus is red. Helices are numbered 1-5.

this system, comprising 5 α-helices, numbered from N to C terminal, and a single antiparallel β-sheet. In this way, the ArsR HTH motif, and hence suspected DNA binding site, incorporates helices α3 and α4 plus the turn between them. This model is based on the NMR structure of NmtR which, incidentally, was not included in the original test set but is a nickel and cobalt binding transcriptional repressor found in *Mycobacterium tuberculosis*, similar to KmtR but with a different binding environment [51].

Figure 2.7 displays the structure and disorder confidence levels associated with the model shown in Figure 2.6. It can be seen that, for the most part, the only regions with high

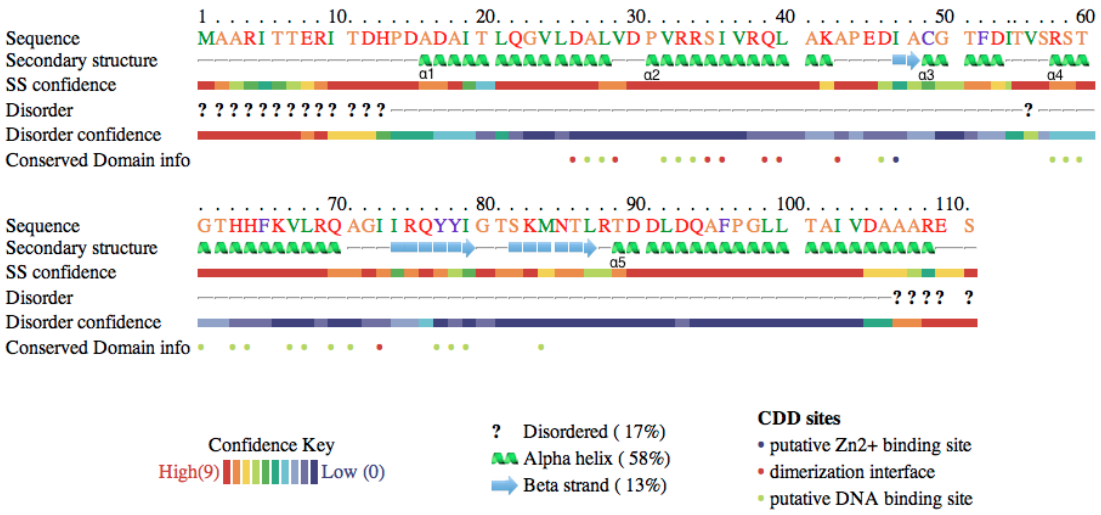


Figure 2.7: Confidence levels in prediction of secondary structure (helices labelled as in Figure 2.6) and local disorder of MmyJ homology model based on template c2lkpB. Residue text colour indicates hydrophobic (green), small/polar (orange), charged (red) or aromatic/cysteine (purple) nature.

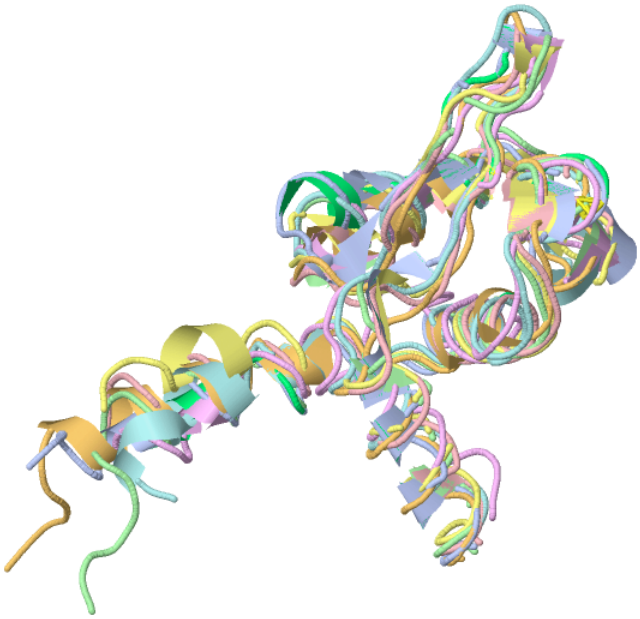


Figure 2.8: Superposition of the c2lkpB model with the 6 other models with a TM-score of 0.7 or higher when compared to c2lkpB. These models were based on templates c1r22B (SmtB), d1r1ta, d1r1ua, c4omzG (NolR), c3jthA (HlyU) and c3cuoB (YgaV). RMSD of all models is between 1.484 and 1.995 Å

disorder/low structure confidence are the ends of the molecule (typically the most flexible part of a protein), with a single point of high disorder within the turn between helices 3 and 4, although as this is the suspected DNA binding site, a degree of flexibility is expected.

Finally, Figure 2.8 displays a superposition of the 6 other models produced by the Phyre2 algorithm with a confidence of 99.9% that are also classified as 'Highly Similar' to the c2lpB model by having a TM-score (measuring the degree of similarity between tertiary protein structures on a scale of 0 to 1 [114]) of 0.7 or higher when compared to this best predicted structure. As expected from the confidence levels of the c2lpB model, it can be seen that the highest level of degeneracy is at the ends of the structure, with good agreement between the models demonstrated for the majority of α-helical residues.

2.3.2 Dimeric Interface

In order to model the expected hydrogen bonded homodimer, the pdb file generated by Phyre2 for MmyJ was uploaded to the HADDOCK web portal [115, 116]. The resulting docking model can be seen in Figure 2.9 and it is apparent that there is significant deviation around the N terminal region from Figure 2.6, with a longer disordered region and helix α1 appearing much shorter. However, as the Phyre model had high confidence of disorder in this region, there is

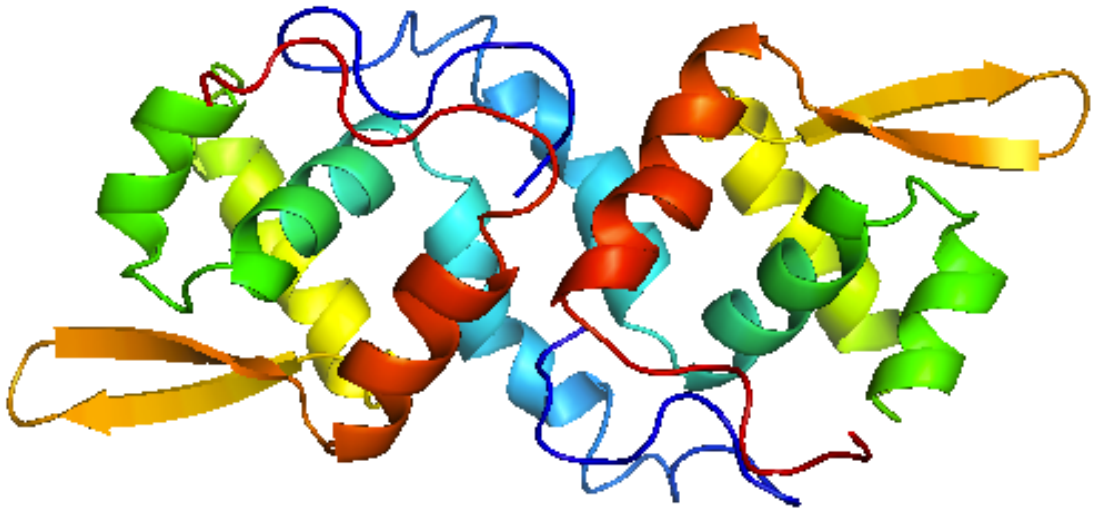


Figure 2.9: Predicted MmyJ dimer modelled by HADDOCK [115, 116]. Each individual chain is coloured blue to red from N to C terminus as in Figure 2.6.

no concern over this issue. The Root Mean Squared Deviation (RMSD) between this predicted dimer model and the Phyre2 monomer was calculated using PyMOL [42] as shown below

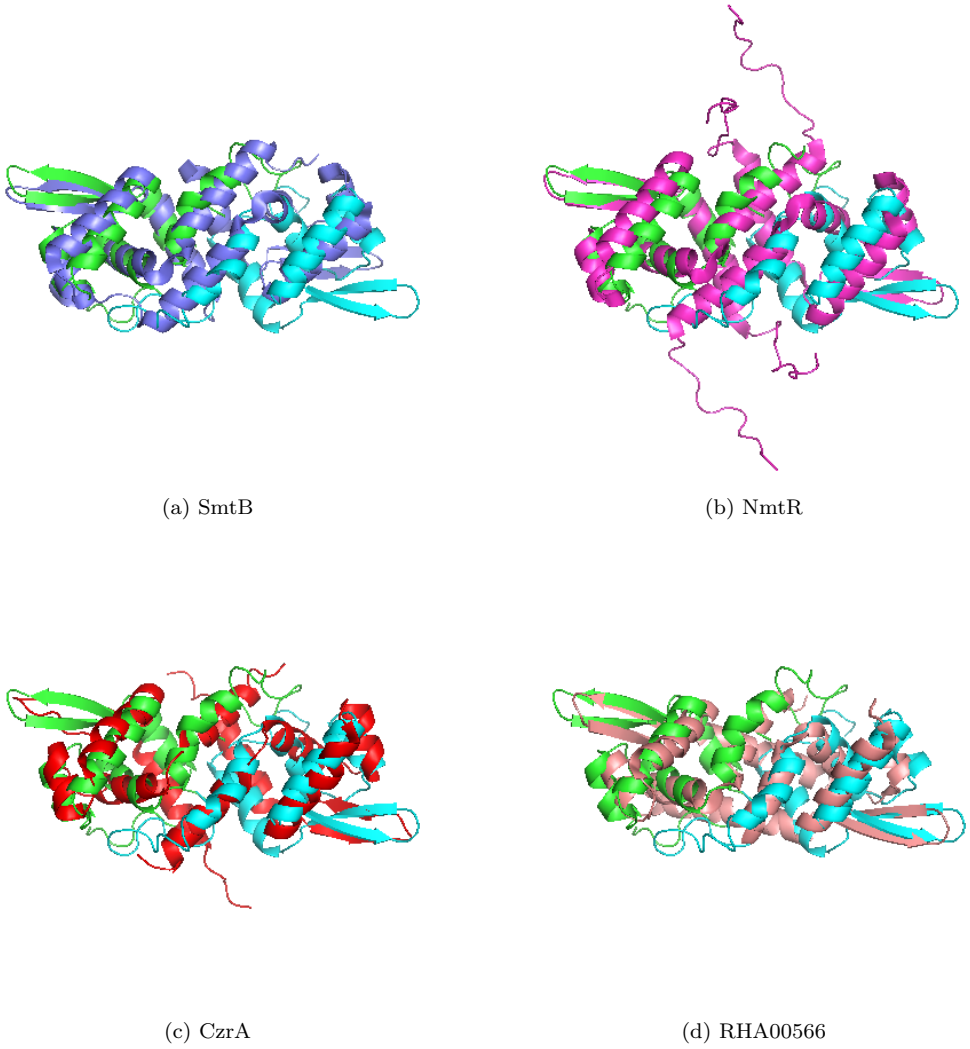
$$RMSD = \sqrt{\frac{\sum {d_{ii}}^2}{N}} \tag{2.1}$$

where d_{ii} is the distance between the i^{th} atoms in each sequence and N is the number of atoms matched in each structure. In this case, the RMSD between the Phyre2 monomeric model and HADDOCK dimeric model was calculated to be 2.132 Å across 88 atoms (excluding outliers). It is thought that this could be a demonstration of the conformational change required to bind to DNA, although structural data would ideally be needed to confirm this assumption.

It is interesting to note that the dimeric interface predicted by HADDOCK matches the approximate interface locations predicted initially by BLAST based on SmtB (see Figure 2.3), with helices 1 and 5 providing the majority of the dimer interface, leaving helices 3 and 4 (thought to be the ArsR HTH domain) exposed on the edge of the dimer for each monomer.

2.3.3 Alignment with ArsR Structures

Following on from the previously mentioned RMSD calculation between the predicted MmyJ monomer and dimer, two structures solved by solution NMR and two solved by X-ray crystallography were obtained for other ArsR family proteins from the PDB so that a similar alignment



(a) SmtB (b) NmtR

(c) CzrA (d) RHA00566

Figure 2.10: Alignments of four solved ArsR family structures with HADDOCK model of MmyJ dimer. In all figures the MmyJ dimer is coloured lime green and turquoise for the two individual monomers, while the compared structure is shown as a single colour.

Name	**PDB ID**	**Structure Type**	**Citation**	**RMSD (wrt MmyJ)**
SmtB	1R1T	XRD (1.70 Å)	[84]	4.542 Å over 139 atoms
NmtR	2LKP	Solution NMR	[117]	4.318 Å over 131 atoms
CzrA	2KJB	Solution NMR	[81]	5.106 Å over 155 atoms
RHA00566	3F6O	XRD (1.90 Å)	[118]	7.339 Å over 136 atoms

Table 2.3: Details of protein structures compared to predicted MmyJ dimer in Figure 2.10. Structures are either characterised by X-Ray Diffraction (XRD) or Nuclear Magnetic Resonance (NMR) experiments. RMSD calculated as shown in Equation 2.1.

could be carried out against those. Again, this was done in Pymol with RMSD values calculated as before. Figure 2.10 illustrates these alignments, with details of the structures used shown in Table 2.3.

While the RMSD values shown in Table 2.3 are higher than those between the solved structures (for example the RMSD between SmtB and NmtR was found to be just 2.285 Å over 178 atoms), this is understandable given that the MmyJ structure is only a predicted model. However, it is still thought that these RMSD values are low enough to indicate a good match, giving confidence in the HADDOCK structure. Also, through visual inspection of the overlay of the solved structures with that predicted of MmyJ, it is clear that the dimeric interface appears to have been predicted correctly and that the ArsR HTH regions are in similar positions, further indicating that this simulation is likely a good prediction of the MmyJ structure.

2.4 MmyJ Orthologues

Methylenomycin-related gene clusters have been identified in two other *Streptomyces* species via NCBI BLAST [101]: *S.* sp. NRRL_S-31 [119] and *S.* sp. 251MFTsu5.1 [120]. Protein sequences encoded within these clusters were previously identified as the two hits from BLASTp analysis with the highest percentage identity to MmyJ and, as such, are thought to be examples of orthologues of MmyJ. When comparing the *mmr* and *mmyJ* genes and associated intergenic region of these other clusters it was noted that the orthologues were both 123 amino acids

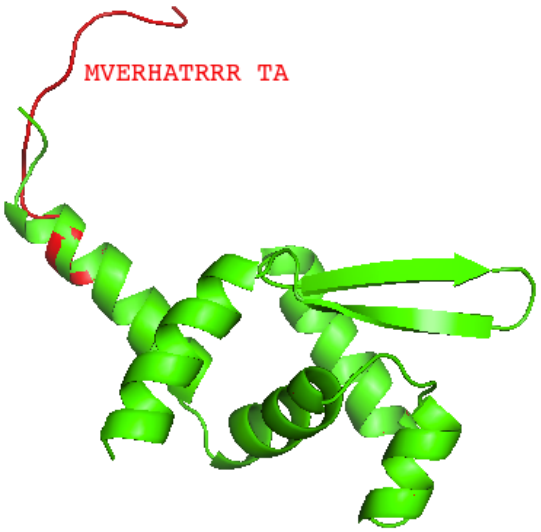


Figure 2.11: Phyre2 models of original MmyJ sequence (green) and N-terminal extended version (red) annotated with the extra 12 amino acids incorporated.

```
MmyJ (ext)   MVERHATRRRTAVAARITTERITDHPDADAITLQGVLDALVDPVRRSIVRQLAKAPEDIA  60
NRRL_S-31    MAGQHSTRRRTTVAARITTERITDHPEADAINLQGVLDALVDPVRRSIVRQLAQAPEDIA  60
351MFTsu5.1  ---MCSGTGRHEVTARTAEERITDHPDAQDITLQGVLEALADPVRRSIVRQIAGAADERK  57
                   :   *  *:** : *******:*: *.*****:**.**********:* *.::

MmyJ (ext)   CGTFDITVSRSTGTHHFKVLRQAGIIRQYYIGTSKMNTLRTDDLDQAFPGLLTAIVDAAA 120
NRRL_S-31    CGTFDITVSRSTGTHHFKVLRQAGIIRQYYVGTSKMNALRIDDLEQAFPGLLTAIVNAAA 120
351MFTsu5.1  CGTFDISVTRSTGTHHFRVLRQSGIIRQYYMGTSKMNILRRDDLDRAFPGLLDAVVEAAN 117
             ******:*:********:****:*******:****** ** ***::****** *:*:**

MmyJ (ext)   RES--- 123
NRRL_S-31    RET--- 123
351MFTsu5.1  REVARR 123
             **
```

Figure 2.12: ClustalW2 alignment of the extended version of MmyJ from *S. coelicolor* A3(2) compared with orthologues from *S.* sp. NRRL_S-31 and *S.* sp. 351MFTsu5.1. Amino acids not previously included in MmyJ analyses are underlined. Asterisks indicate fully conserved residues, colons indicate conservation between groups of strongly similar properties and periods indicate conservation between groups of weakly similar properties. Strongly similar and weakly similar are defined as scoring greater than, or less than or equal to 0.5 in the Gonnet PAM 250 matrix [121]. Colours indicate grouping of amino acids as small or hydrophobic (red), acidic (blue), basic (pink) or containing a hydroxyl, sulfhydryl or amino side chain (green).

long; 12 amino acids longer than MmyJ. After further inspection of the *S. coelicolor* A3(2) methylenomycin gene cluster, it was noticed that there was a second possible start codon for MmyJ, which would make it 123 amino acids long as in the orthologous systems. Figure 2.11 shows the possible extended version of MmyJ, along with a new Phyre2 homology model with the extra 12 N-terminal amino acids included. It can be seen that these extra amino acids are predicted to have no impact on the overall fold of the protein, with the only differences being apparent in helix 1, which is now slightly shorter, and the disordered N-terminal region, which is now longer. As such, it is not thought to be important whether these extra amino acids are included, and so all previous analyses are still expected to be valid.

Figure 2.12 shows the protein sequence of MmyJ from *S. coelicolor* A3(2) aligned with the two orthologues from *S.* sp. NRRL_S-31 and *S.* sp. 251MFTsu5.1 using ClustalW2 [122]. It can be seen that there is a high degree of similarity between the three orthologues, with 65% identity across all three amino acid chains. Including those amino acids deemed to have strongly similar properties, the three orthologues can be said to have 80% similarity in total, strongly suggesting that they are indeed true orthologues.

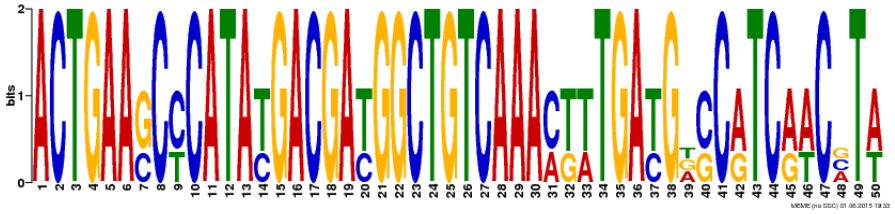
It is interesting to note that the ArsR HTH DNA binding domain, previously identified in MmyJ by PROSITE (transposed now to amino acids 59 to 82 due to the additional 12 amino

acids at the N terminus), is one of the highest conserved regions of the three protein sequences, with 79% identity and 92% similarity across all three orthologues. The region contains only 2 residues not deemed to be at least strongly similar according to the Gonnet PAM 250 similarity matrix [121]. This leads to the expectation that, should MmyJ truly behave as an ArsR protein as thought and inhibit the expression of the efflux pump protein Mmr until required by binding to the DNA between the *mmr* and *mmyJ* genes, then it can be assumed that the DNA target region should also be highly conserved across all three orthologues.

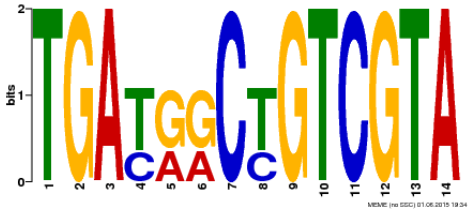
2.5 Target DNA Sequence

Under the assumption that the DNA target region for MmyJ lies in the intergenic region between *mmr* and *mmyJ*, and that it is conserved across all three orthologues, the three intergenic regions were submitted to MEME for analysis, with motifs discovered shown in Figure 2.13 [111]. While the blue and green motifs have e-values higher than 0.05, and can therefore be dismissed as insignificant, the red motif with and e-value of 3.8e-11 can be assumed to have quite high significance and, therefore, it is likely that the DNA binding target lies in this region.

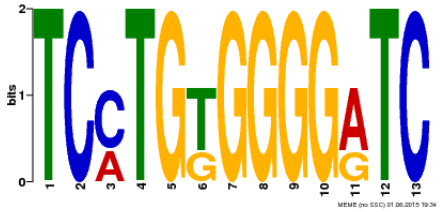
(a)



(b) Red motif; e-value 3.8e-11

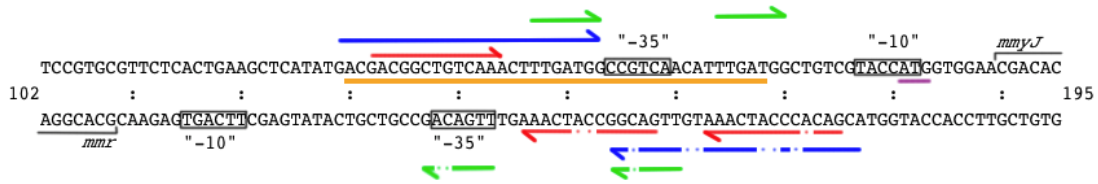


(c) Blue motif; e-value 9.1e+1

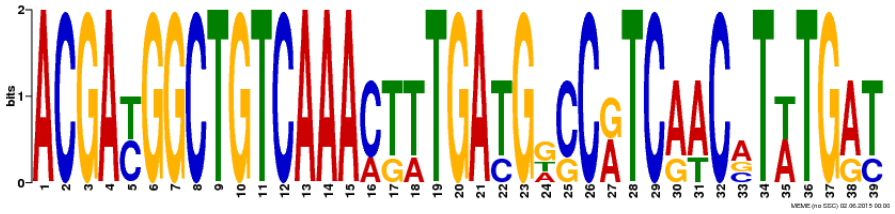


(d) Green motif; e-value 4.8e+2

Figure 2.13: Results of MEME analysis of intergenic region between *mmr* and *mmyJ* genes in *S. coelicolor* A3(2) and analogous systems. Motifs highlighted in (a) are expanded and labelled by colour as motifs (b) - (d).



(a)



(b)

Figure 2.14: (a) Sets of semi-conserved inverted palindromic sequences identified as protected via DNA fingerprinting [100]. *mmyJ* and *mmr* signify the start of the transcription start sites for each gene. Base numbers correspond to 1 being the start of the intergenic region with colons indicating groups of 10 nucleotides from this start point. Sequence reversed compared to Figure 1.16. Purple line identifies start codon of possible extended 123 residue version of MmyJ. Orange line corresponds to motif shown in (b), identified by MEME as being common across all three *mmr-mmyJ* orthologues when weighted towards the semi-conserved palindromic regions identified in [100]. This motif was returned with an e-value of 5.7e-19.

DNA fingerprinting techniques have previously identified a protected area of the *mmr-mmyJ* intergenic region [100] containing a semi-conserved inverted palindromic repeat sequence of the type usually associated with DNA-binding proteins [18, 123]. This region is illustrated in Figure 2.14a. It can be seen that there are several possible sets of semi-conserved inverted palindromic repeats to which MmyJ could bind, all of which are around the -35 promoter sites for both *mmr* and *mmyJ*. This implies that binding at any of these palindromic sites would block the promoter region and, therefore, inhibit production of both MmyJ and Mmr proteins.

It is worth noting at this point that an investigation of this region of the gene cluster was undertaken to see if there was any significance in the placement of the two possible start codons for MmyJ with regard to the promoter site. In this way it was hoped that it could be confirmed whether MmyJ is comprised of the 111 amino acids originally thought, or whether the 123 amino acid extended version is the true sequence. It was found that the ATG start point for the extended version lies within the previously identified -10 site in Figure 2.14a, as shown by sequence underlined in purple. As the origin of transcription of MmyJ has previously been confirmed [99], it can be concluded that the extra 12 amino acids are not incorporated into the protein, and as such that MmyJ is indeed 111 residues long as originally thought.

In order to further investigate these DNA binding sites, the previous MEME analysis was re-run with the total possible palindromic region (as indicated by the blue arrows in Figure 2.14a) added as a fourth sequence, thereby weighting the analysis towards this region. In this way, it was hoped that a motif would be recognised that incorporates one whole palindromic sequence, be it the blue, red or green one indicated in Figure 2.14a. This analysis produced a similar motif pattern to that shown in Figure 2.13, but the main motif recognised is shorter than previously, with a lower e-value of 5.7e-19. This new motif is shown in Figure 2.14b [111] and corresponds to bases 130 to 168 of the *mmr-mmyJ* intergentic region in *S. coelicolor* A3(2). From the regions shown in Figure 2.14a, this can be seen to incorporate either the red palindromic sequence (with half of the extra repeat) or the inner-most green palindromic sequence. As ArsR family proteins typically bind to an imperfect 12-2-12 inverted sequence [28], and the red palindromic sequence highlighted in Figure 2.14a and incorporated into the motif shown in Figure 2.14b is a 13-1-13 semi-conserved inverted palindromic repeat, it is assumed that this is the most likely DNA binding site and will be investigated further by experimentation (see Chapter 5).

3 Protein Overproduction & Purification

3.1 Protein Overproduction

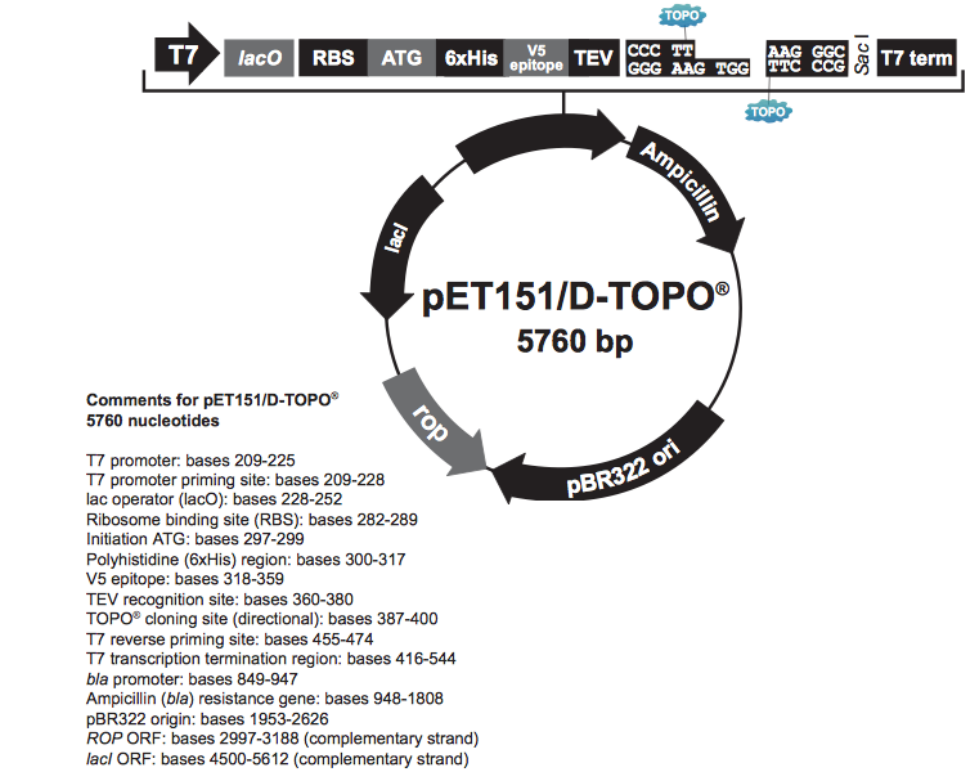
In order to overproduce and purify His_6-MmyJ from *E. coli*, the *mmyJ* gene was first cloned into a suitable expression vector with an inducible promoter.

3.1.1 Cloning *mmyJ*

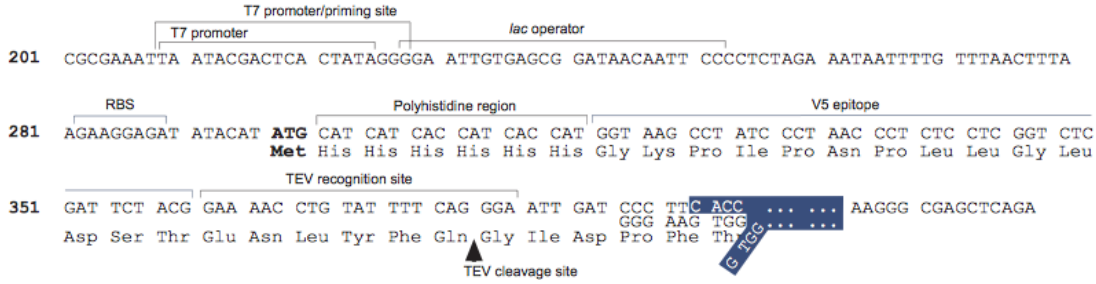
There were several factors to consider when choosing an appropriate expression vector. Firstly, the vector would need to have a fusion tag included so that the recombinant protein could easily be purified. Due to the high affinity of histidine for metal ions, including nickel (see Section 3.1.3), it was decided that a histidine tag would be an appropriate choice. The Champion™ pET Directional TOPO® Expression Kit from Invitrogen™ [124] was chosen due to the ease of using the TOPO cloning method to insert the desired gene into the expression vector. As the vectors in these kits offer a choice of either 6xHis or His-Patch (HP) Thioredoxin tag, consideration had to be given to the type of tag used. Due to the *mmyJ* gene being only 336 bp long, the 6xHis tag (96 bp) was chosen over HP Thioredoxin (330 bp) so that the overall increase in length of the complete fusion protein compared to MmyJ was minimised, in order to reduce the risk of non-native protein folding.

Secondly, it was decided that a cleavage site should be included to allow the removal of the histidine tag, if required, for further experimentation. Again, the range of TOPO cloning kits offered a choice between enterokinase or tobacco etch virus (TEV) cleavage sites. As TEV was already used by another group in the laboratory, and so would be easily available, it was decided that this would be the more suitable choice. This combination of 6xHis tag and TEV cleavage site meant that the pET151/D-TOPO™ plasmid would be a sensible choice for this work.

The plasmid map from the kit's manual is shown in Figure 3.1, with the inclusion of an ampicillin resistance gene that can be used as a selection marker. It can be seen, that as well as the 6xHis tag, a V5 epitope is included which, together with the TEV cleavage site, brings the total tag length to 96 bp, corresponding to approximately a third of the length of the *mmyJ* gene. After TEV protease cleavage there would only be 6 amino acids extra added to the



(a)



(b)

Figure 3.1: (a) pET151 plasmid map and (b) DNA sequence of 6xHis tag (and corresponding amino acid sequence) including preceeding *lac* operator, taken from [124]. Complete vector sequence is available from `http://www.invitrogen.com`.

N-terminus of the recombinant MmyJ, which was expected to have a negligible effect on the protein fold, even if there was an effect from the full length tag.

The *mmyJ* gene was amplified for insertion into the pET151 vector by PCR using Roche High Fidelity Polymerase and oligonucleotide pair 2 (see Section 8.1.3). The resulting PCR product was inserted into the pET151 plasmid in accordance with the instructions in [124], which was then transformed into *E. coli* Top10 cells. A colony was picked and a glycerol stock was prepared. The plasmid was extracted from a culture inoculated with this glycerol stock and 20 µL was sent for sequencing by GATC with the standard T7 primer (oligonucleotide 1). Figure 3.2 shows the resulting sequencing data, along with the expected sequence for the pET151

```
pET151mmyJ            1 --------------------------------------------------atg
MmyJ-Top10-T7         1 gaattcctctagaatattttgtttaactttaagaaggagatatacatatg

pET151mmyJ            4 catcatcaccatcaccatggtaagcctatccctaaccctctcctcggtct
MmyJ-Top10-T7        51 catcatcaccatcaccatggtaagcctatccctaaccctctcctcggtct

pET151mmyJ           54 cgattctacggaaaacctgtattttcagggaattgatcccttcaccgtgg
MmyJ-Top10-T7       101 cgattctacggaaaacctgtattttcagggaattgatcccttcaccgtgg

pET151mmyJ          104 cggcacggatcacgacagagcgcatcaccgaccatccggacgctgacgcc
MmyJ-Top10-T7       151 cggcacggatcacgacagagcgcatcaccgaccatccggacgctgacgcc

pET151mmyJ          154 atcaccctccagggcgtcctggacgcgctggtcgatccggtgcgccgcag
MmyJ-Top10-T7       201 atcaccctccagggcgtcctggacgcgctggtcgatccggtgcgccgcag

pET151mmyJ          204 catcgtccggcagctggctaaggcacccgaggacatcgcctgcggcacct
MmyJ-Top10-T7       251 catcgtccggcagctggctaaggcacccgaggacatcgcctgcggcacct

pET151mmyJ          254 tcgacatcaccgtctcccgctcgaccggcactcaccacttcaaggtgttg
MmyJ-Top10-T7       301 tcgacatcaccgtctcccgctcgaccggcactcaccacttcaaggtgttg

pET151mmyJ          304 cgccaggccgggatcatcaggcagtactacatcggcacctcgaagatgaa
MmyJ-Top10-T7       351 cgccaggccgggatcatcaggcagtactacatcggcacctcgaagatgaa

pET151mmyJ          354 cacgcttcgcaccgatgatctcgatcaggccttccccggcctgctcaccg
MmyJ-Top10-T7       401 cacgcttcgcaccgatgatctcgatcaggccttccccggcctgctcaccg

pET151mmyJ          404 cgatcgtcgacgccgcggccagggagagctgaccggccaccgctcgcccg
MmyJ-Top10-T7       451 cgatcgtcgacgccgcggccagggagagctgaccggccaccgctcgcccg

pET151mmyJ          454 cacggcgctccaaagggcgagctcagatccggctgctaacaaagcccgaa
MmyJ-Top10-T7       501 cacggcgctccaaagggcgagctcagatccggctgctaacaaagcccgaa

pET151mmyJ          504 aggaagctgagttggctgctgccaccgctgagcaataactagcataaccc
MmyJ-Top10-T7       551 aggaagctgagttggctgctgccaccgctgagcaataactagcataaccc

pET151mmyJ          554 cttggggcctctaaacgggtcttgaggggttttttgctgaaaggaggaac
MmyJ-Top10-T7       601 cttggggcctctaaacgggtcttgaggggttttttgctgaaaggaggaac

pET151mmyJ          604 tatatccggatatcccgcaagaggcccggcagtaccggcataaccaagcc
MmyJ-Top10-T7       651 tatatccggatatcccgcaagaggcccggcagtaccggcataaccaagcc

pET151mmyJ          654 tatgcctacagcatccagggtgacggtgccgaggatgacgatgagcgcat
MmyJ-Top10-T7       701 tatgcctacagcatccagggtgacggtgccgaggatgacgatgagcgcat

pET151mmyJ          704 tgttagatttcatacacggtgcctgactgcgttagcaatttaactgtgat
MmyJ-Top10-T7       751 tgttagatttcatacacggtgcctgactgcgttagcaatttaactgtgat

pET151mmyJ          754 aaactaccgcattaaagctagcttatcgatgataagctgtcaaacatgag
MmyJ-Top10-T7       801 aaactaccgcattaaagctagcttatcgatgataagctgtcaaacatgag

pET151mmyJ          804 aattaattcttgaagacgaaagggcctcgtgatacgcctatttttatagg
MmyJ-Top10-T7       851 aattaattcttgaagacgaaagggcctcgtgatacgcctatttttatagg
```

(Figure Continued Overleaf)

```
pET151mmyJ          854 ttaatgtcatgataataatggtttcttagacgtcaggtggcacttttcgg
MmyJ-Top10-T7       901 ttaatgtcatgataataatggtttcttagacgtcaggtggcacttttcgg

pET151mmyJ          904 ggaaatgtgcgcggaacccctatttgtttatttttctaaatacattcaaa
MmyJ-Top10-T7       951 ggaaatgtgcgcggaacccctatttgtttatttttctaaatacattcaaa

pET151mmyJ          954 tatgtatccgctcatgagacaataaccctgataaatgcttcaataatatt
MmyJ-Top10-T7      1001 tatgtatccgctcatgagacaataaccctgataaatgcttcaataatatt

pET151mmyJ         1004 gaaaaaggaagagtatgagtattcaacatttccgtgtcgcccttattccc
MmyJ-Top10-T7      1051 gaaaaaggaagagtatgagtattcaacatttccgtgtcgcccttattccc

pET151mmyJ         1054 ttttttgcggcattttgccttcctgtttttgctcacccagaaacgctggt
MmyJ-Top10-T7      1101 ttttttgcggcattttgccttcctgtttttgctcacccagaaacgctggt

pET151mmyJ         1104 gaaagtaaaagatgctgaagatcagttgggtgcacgagtgggttacatcg
MmyJ-Top10-T7      1151 gaaagtaaa-----------------------------------------

pET151mmyJ         1154 aactggatctcaacagcggtaagatccttgagagttttcgccccgaagaa
MmyJ-Top10-T7           ----------------------------------------------------

pET151mmyJ         1204 cgttttccaatgatgagcacttttaaagttctgctatgtggcgcggtatt
MmyJ-Top10-T7           ----------------------------------------------------

pET151mmyJ         1254 atcccgtgttgacgccgggcaagagcaactcggtcgccgcatacactatt
MmyJ-Top10-T7           ----------------------------------------------------

pET151mmyJ         1304 ctcagaatgacttggttgagtactcaccagtcacagaaaagcatcttacg
MmyJ-Top10-T7           ----------------------------------------------------

pET151mmyJ         1354 gatggcatgacagtaagagaattatgcagtgctgccataaccatgagtga
MmyJ-Top10-T7           ----------------------------------------------------

pET151mmyJ         1404 taacactgcggccaacttacttctgacaacgatcggaggaccgaaggagc
MmyJ-Top10-T7           ----------------------------------------------------

pET151mmyJ         1454 taaccgctttttttgcacaacatgggggatcatgtaactcgccttgatcgt
MmyJ-Top10-T7           ----------------------------------------------------

pET151mmyJ         1504 tgggaaccggagctgaatgaagccataccaaacgacgagcgtgacaccac
MmyJ-Top10-T7           ----------------------------------------------------

pET151mmyJ         1554 gatgcctgcagcaatggcaacaacgttgcgcaaactattaactggcgaac
MmyJ-Top10-T7           ----------------------------------------------------

pET151mmyJ         1604 tacttactctagcttcccggcaacaattaatagactggatggaggcggat
MmyJ-Top10-T7           ----------------------------------------------------
```

Figure 3.2: Sequenced pET151-mmyJ plasmid. Inserted *mmyJ* gene runs from base 100 to 435 inclusive. 'pET151mmyJ' is the designed construct, and 'MmyJ-Top10-T7' is the result of sequencing the constructed plasmid extracted from transformed Top10 cells. Yellow highlight indicates agreement between sequences. Only the first 1653 bases of the pET151 plasmid are shown here.

plasmid with *mmyJ* gene inserted. As can be seen, the sequenced plasmid and predicted construct align perfectly with the inserted gene sequenced in its entirety. It can therefore be concluded that the *mmyJ* gene had been successfully cloned into the pET151/D-TOPO plasmid.

3.1.2 Overproduction of MmyJ

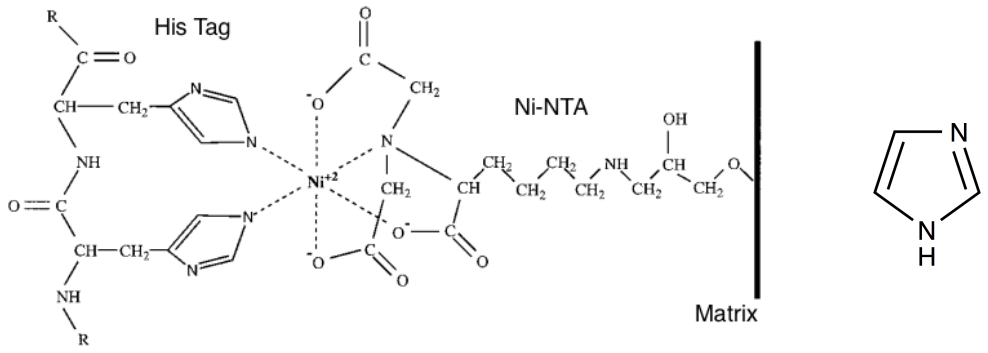
BL21 star cells carry the DE3 bacteriophage λ lysogen containing the gene for T7 RNA polymerase, required for transcription of the cloned gene in pET plasmids. This is under the control of a *lac* promoter, meaning that, in theory, no T7 RNA polymerase is produced by the cells until IPTG induction. However, studies have shown that the λDE3 lysogen can produce small amounts of T7 RNA polymerase even without IPTG present [125]. In order to prevent premature expression of proteins that may hinder the growth of the BL21 star cell culture, a second *lac* operator is included in pET TOPO plasmids downstream of the T7 promoter, adding a second degree of regulation to the expression system when in BL21 star [13, 12].

The pET151-*mmyJ* plasmid was transformed into BL21 star cells and the 6xHis tagged MmyJ, from here on referred to as His_6-MmyJ, was expressed. As BL21 star is known to reduce in efficiency over time due to homologous recombination by RecA [126], fresh BL21 star cells were regularly transformed with pET151-*mmyJ* plasmid to maintain optimal expression levels when used to inoculate cultures.

3.1.3 Chromatographic Purification

In order to purify His_6-MmyJ from the Cell Free Extract (CFE), a Fast Protein Liquid Chromatography (FPLC) instrument was used to perform Immobilised Metal Ion Chromatography (IMAC).

FPLC works by drawing from one or more reservoirs of buffer in order to flow proteins in solution (liquid phase) through a column containing an immobilised resin (stationary phase) [127]. Buffer conditions can then be altered so as to separate proteins in the column either by size/shape, charge, hydrophobicity, function, or interaction with ligands [128]. Thus, proteins can be purified by, for example, size exclusion chromatography, immobilised metal affinity chromatography (IMAC) or anion/cation exchange through the use of specifically designed



(a) (b)

Figure 3.3: (a) Demonstration of binding of polyhistidine tag to the Ni^{2+} ion in nickel-nitrilotriacetic acid (Ni-NTA) molecule immobilised on a matrix, modified from [135]. (b) Structure of imidazole, which competitively binds to the Ni^{2+} over the polyhistidine tag, causing elution of the immobilised protein.

pre-packed columns containing micron-scale gel beads with varying properties [129]. FPLC varies only slightly from High Pressure Liquid Chromatography (HPLC), with the main differences being that FPLC operates at lower pressure and is capable of running larger samples ranging from 100 µL to 150 mL or more, whereas a typical HPLC system handles samples of 100 µL at most. Also, FPLC systems have simpler detectors, typically only operating at wavelengths relevant to protein purification (i.e. 254 nm and 280 nm) [130]. This makes FPLC a valuable tool in the bulk purification of expressed proteins from a large liquid culture, easily able to purify upwards of 10 mg of protein from 50 mL of CFE.

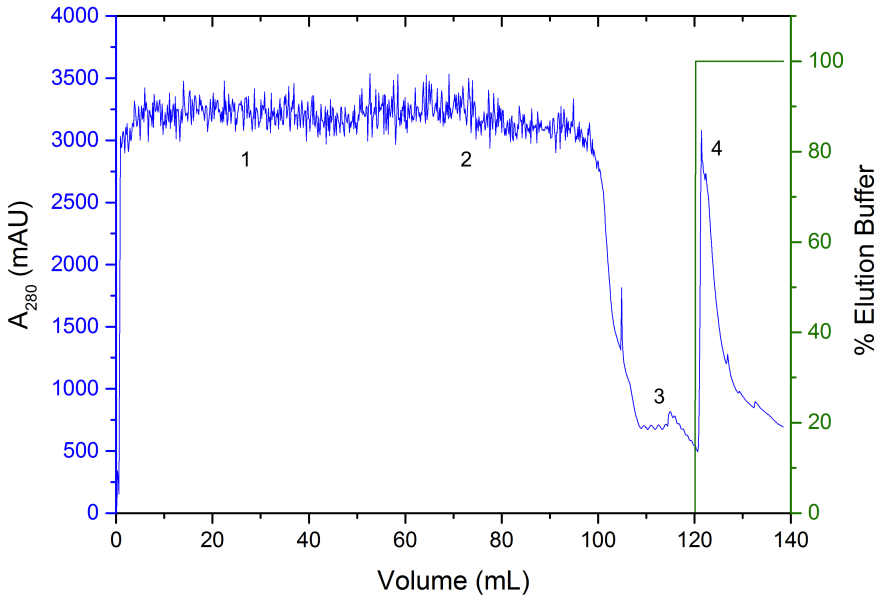
IMAC relies on the interactions between metal ions immobilised on a chelating ligand and amino acid residues, specifically histidine, tryptophan and cysteine [131], and was first proposed in 1975 under the name Metal-Chelate Affinity Chromatography [132]. The chelating ligand used can be optimised for different types of tags or antibodies, leading to a breadth of purification options [133]. Once bound to the chelating ligand, the tagged protein can then be eluted by competitive binding with certain salts or small molecules, depending on the chelating ligand used. In all IMAC purifications carried out here, the chelating ligand was nitrilotriacetic acid (NTA) bound to Ni^{2+}, with imidazole being used to elute the purified protein by competitively binding to the nickel ions [134], as detailed in Figure 3.3.

It has already been mentioned that the pET plasmid was selected with metal affinity chromatography in mind, specifically the affinity between histidine and Ni^{2+}, allowing IMAC to be performed. Various elution methods were trialled so as to best separate weakly binding

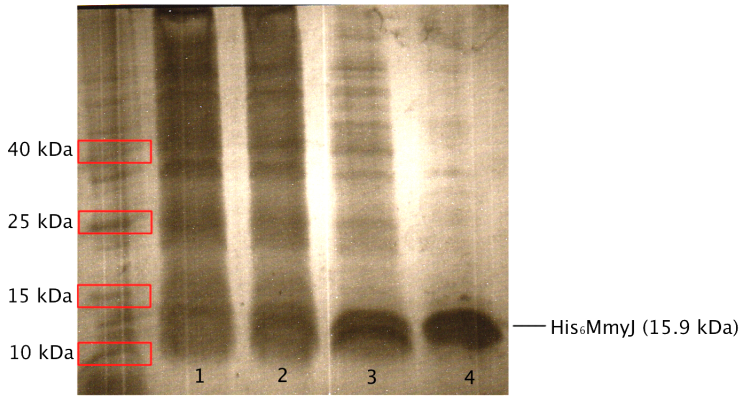
contaminants such as the *E. coli* protein SlyD, a 29 kDa protein with a high natural affinity to nickel [136] which could cause false results in later experiments due to the similarity between its mass and that of potential His_6-MmyJ dimers. Sodium Dodecyl Sulphate Poly-Acrylamide Gel Electrophoresis (SDS-PAGE) was used to observe the proteins in chosen fractions from each FPLC purification.

3.1.4 Optimising Protein Elution from Ni^{2+} Columns

The first elution method attempted was a simple switch between wash/binding buffer and elution buffer. These first experiments also had 20 mM imidazole added to the wash/binding



(a)



(b)

Figure 3.4: (a) FPLC UV absorption trace at 280 nm of His_6-MmyJ purification with corresponding % elution buffer, using wash/binding buffer with 20 mM imidazole and eluting in one step. (b) 15% SDS-PAGE gel showing proteins present in collected fractions. Lane numbers correspond to volumes at positions indicated by numbers on UV trace.

buffer in order to prevent non-specific binding by contaminants such as SlyD, in accordance with the manual supplied with the HisTrap columns [137].

The FPLC trace for this purification, as well as the SDS-PAGE gel of fractions taken, can be seen in Figure 3.4. This purification was done with a larger than normal CFE: approximately 100 mL, and so the experiment was paused after 50 mL so that the Superloop™ could be refilled. Hence the gel has two lanes corresponding to the non-binding proteins; one from each run of the Superloop™ for completeness. It is immediately apparent that the overproduced protein is present in the elution lanes, meaning that it was present in the CFE. As all membrane proteins had previously been removed from the CFE by pelleting by centrifuge, it can therefore be concluded that MmyJ is soluble[6]. However, eluting in this manner has retained some impurities in the purified fraction, as indicated by faint bands higher up in lane 4 in Figure 3.4b (in which there is slight leakage into the ladder from the lane 1). Also, a noticeable amount of overproduced protein is eluted with the non-binding proteins (lanes 1 and 2) and when washing the column to remove any last impurities from it before elution (lane 3).

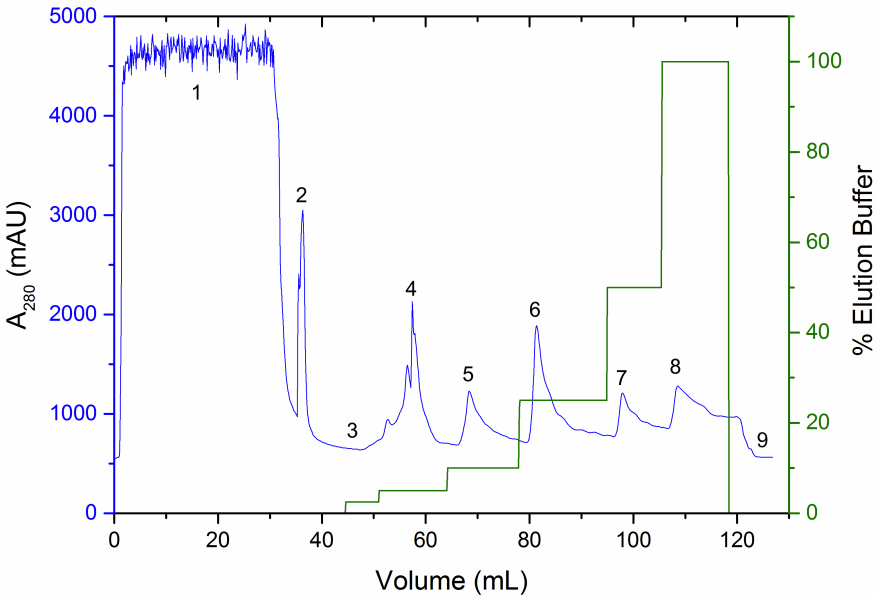
It is worth noting that, as demonstrated in Figure 3.4b, the band for His_6-MmyJ was consistently lower than expected for all gels run. Mass spectrometry results (see Section 3.3) confirm that the bands do correspond to MmyJ, and so it must be concluded that the deviation from the ladder is due to the shape of the protein affecting how it migrates through the gel.

In an attempt to maximise the yield of overproduced His_6-MmyJ, the next stragegy attempted was to reduce the amount of imidazole in the wash/binding buffer to 10 mM and to elute by stepwise increments in percentage of elution buffer. In this way the actual concentration of imidazole in mM can be calculated as:

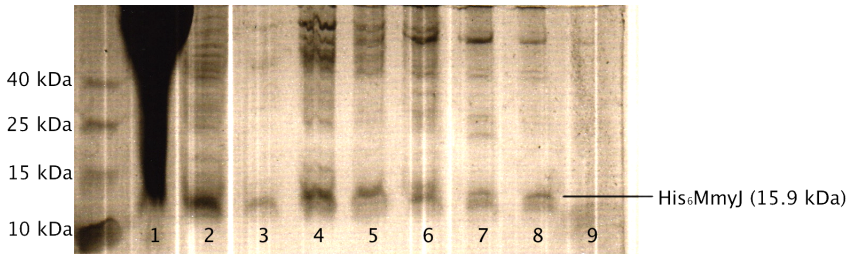
$$[\text{Imidazole}] = (10 \times (100 - \%\ \text{Elution Buffer}) + 200 \times (\%\ \text{Elution Buffer})) \div 100 \qquad (3.1)$$

The results of this run can be seen in Figure 3.5, showing that there are still significant impurity bands in the eluted fractions. Figure 3.5b also indicates that reducing the imidazole content of the wash/binding buffer to 10 mM still leads to partial elution of His_6-MmyJ when washing the column as there is a band present in lane 2. It is also worth noting that this strategy led to

[6] If this was not the case then further work to extract the protein from the pelleted cell debris and insoluble proteins after lysis would be required.



(a)



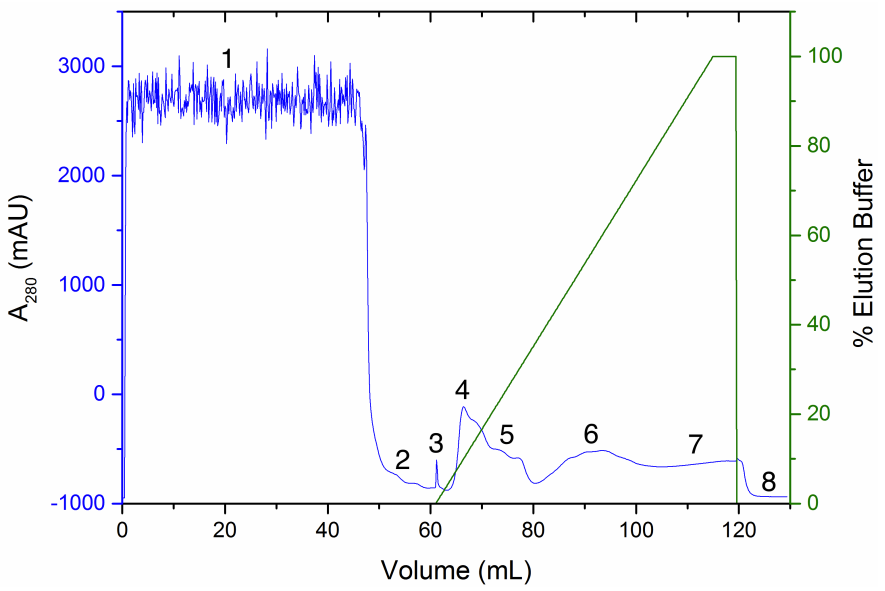
(b)

Figure 3.5: (a) FPLC UV absorption trace at 280 nm of His_6-MmyJ purification with corresponding % elution buffer, using wash/binding buffer with 10 mM imidazole and eluting in several steps. (b) 15% SDS-PAGE gel showing proteins present in collected fractions. Lane numbers correspond to volumes at positions indicated by numbers on UV trace.

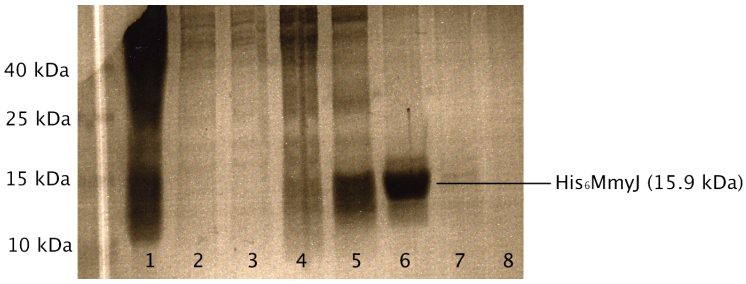
excessive dilution of the eluted protein, as it was present in each of the peaks in varying degrees, and so the fractions needed re-combining and concentrating in order to carry out further work. Finally, there is a distinct drop in absorption when elution is halted by switching back to 100% wash/binding buffer, corresponding to the absorption at 280 nm by imidazole.

The third attempt at optimising the elution protocol is shown in Figure 3.6. In this instance the elution was carried out by gradually increasing the percentage of elution buffer from 0 to 100% over the course of two hours. Also, there was no imidazole present in the wash/binding buffer, meaning the concentration of imidazole in Figure 3.6a can simply be found by multiplying the percentage elution buffer by 2.

It can be seen from Figure 3.6a that the elution profile is more complex in this instance, hence the comparatively large number of samples taken around the largest peak in order to determine



(a)



(b)

Figure 3.6: (a) FPLC UV absorption trace at 280 nm of His_6-MmyJ purification with corresponding % elution buffer, using wash/binding buffer without imidazole and eluting continuously over 2 hours. (b) 15% SDS-PAGE gel showing proteins present in collected fractions. Lane numbers correspond to volumes at positions indicated by numbers on UV trace.

the proteins present via SDS-PAGE. The results seen in this gel (Figure 3.6b) surprisingly show that the largest peak, corresponding to lane 4, contains very little of the overproduced His_6-MmyJ. This peak at low imidazole levels corresponds to the bulk of protein impurities that bound non-specifically to the HisTrap column, whereas His_6-MmyJ does not start to elute until the shoulder of this largest peak, as shown in lane 5, which as such still contains several impurities. However, lane 6 shows very pure protein eluted at higher imidazole concentrations, with the corresponding peak being over a smaller volume than the collection of peaks seen when eluting in a step-wise fashion. It was decided that, even though a large proportion of His_6-MmyJ eluted alongside impurities, this could be kept and re-purified if needed. Also, the second elution peak containing pure His_6-MmyJ was of a higher purity using this elution

method than the others attempted. As such, all His_6-MmyJ was purified using this method.

It should be noted that many elution peaks corresponding to His_6-MmyJ seem small compared to the amount of protein evident in the associated SDS-PAGE gels. This is due to the absence of tryptophan residues in MmyJ, drastically reducing the amount of absorption at 280 nm. It can be thought that a small peak can indicate pure protein, as even a small amount of impurities containing tryptophan would lead to much higher absorption peaks.

3.1.5 TEV Protease Expression and Cleavage of MmyJ Tag

Rosetta (DE3) pLysS cells containing an expression vector for His_6-tagged TEV solubilised through fusion with superfolded green fluorescent protein (sGFP) were obtained from the Dixon Group, who had previously obtained the cells line from the Moreman group at the Complex Carbohydrate Research Centre, University of Georgia (USA). Despite the advantages of BL21 star over Rosetta (DE3) pLysS cells, which do not contain a second *lac* operator regulating T7 RNA polymerase production, it was decided to keep the plasmid in this strain. This was due to previously reported overproduction levels of the sGFP-TEV-His_6 fusion protein in this system being more than adequate for the requirements of these experiments [138, 139]. The expression vector can be seen in Figure 3.7.

An 800 mL LB culture of the Rosetta cells containing this sGFP-TEV-His_6 expression vector was grown with protein expression and purification carried out as previous. The FPLC trace

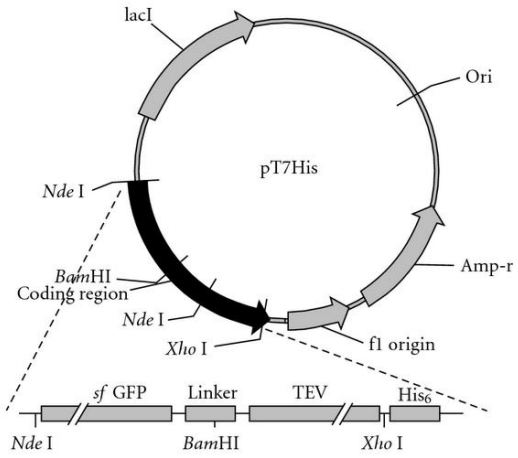
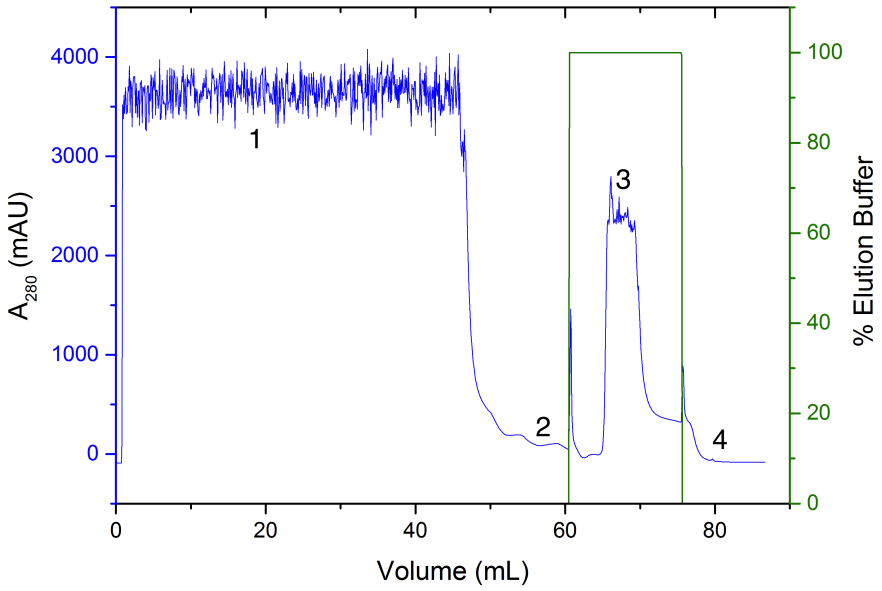
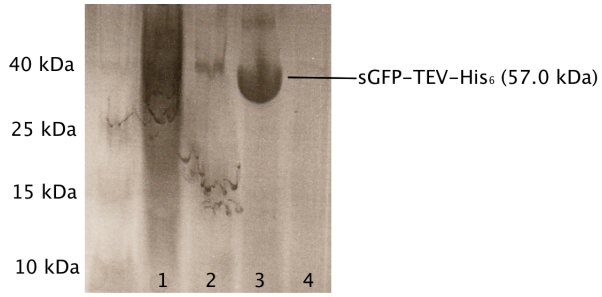


Figure 3.7: Expression vector for sGFP-TEV-His_6, taken from [139]. Linker sequence between sGFP and TEV is defined as GSKGP. Plasmid incorporates ampicillin resistance marker and IPTG inducible expression system.



(a)

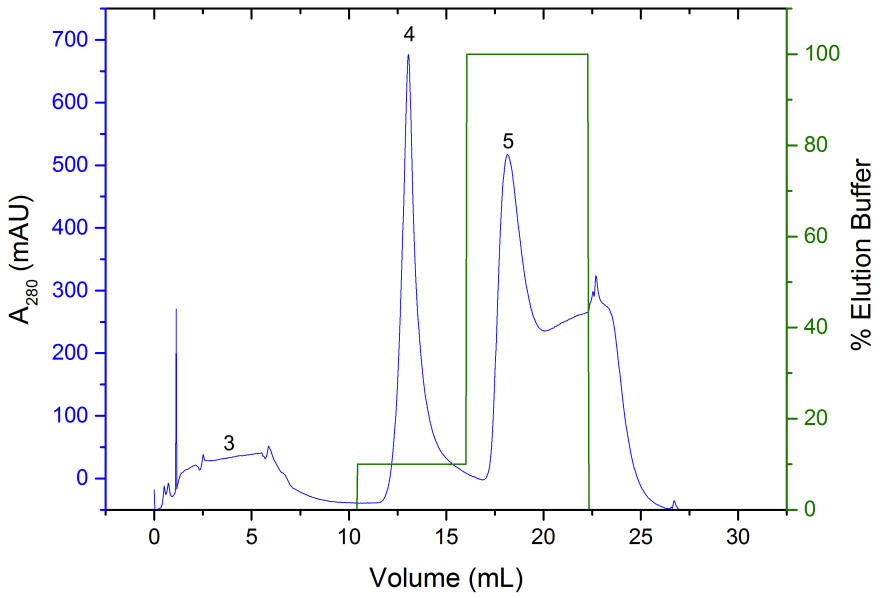


(b)

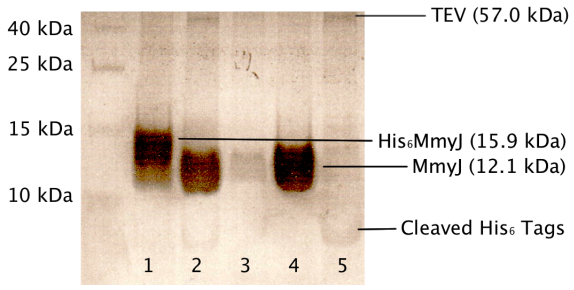
Figure 3.8: (a) FPLC UV absorption trace at 280 nm of TEV purification with corresponding % elution buffer, using wash/binding buffer without imidazole and eluting in a single step. (b) 15% SDS-PAGE gel showing proteins present in collected fractions. Lane numbers correspond to volumes at positions indicated by numbers on UV trace.

at 280 nm for the purified TEV, with accompanying SDS-PAGE gel, can be seen in Figure 3.8. A simple one step elution was used in this instance as it was deemed that any other protein impurities retained would be removed along with the TEV and 6xHis tags after cleavage. It is worth noting that, as with His_6-MmyJ, there is a discrepancy in Figure 3.8b between the indicated molecular weight of the purified protein and the ladder. In this instance it was easy to see that the band was indeed sGFP-TEV-His_6 as the band was visibly green before staining. The eluted TEV was then stored at 4°C.

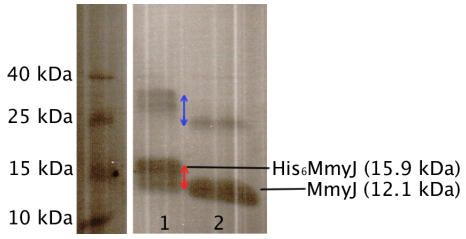
His tag cleavage was carried out according to [140]. IMAC FPLC was then performed on the sample again, as shown in Figure 3.9, with some cleaved MmyJ passing through the column (demonstrated by a weak band in lane 3 in Figure 3.9b), although the majority of the cleaved protein was retained by the column due to non-specific binding, along with the TEV



(a)



(b)



(c)

Figure 3.9: (a) FPLC UV absorption trace at 280 nm of purification of MmyJ from cleaved 6xHis tags and sGFP-TEV-His_6 protease, using wash/binding buffer with no imidazole. Cleaved MmyJ was eluted using 20 mM imidazole (10% elution buffer) due to non-specific binding to the His-Trap column. (b) 15% SDS-PAGE gel showing proteins present in fractions after TEV cleavage. Lanes 1 and 2 correspond to His_6-MmyJ samples before and after TEV was added, demonstrating reduction in molecular weight due to loss of tags, with further lane numbers corresponding to volumes at positions indicated on UV trace. (c) 15% SDS-PAGE gel showing another purification of cleaved MmyJ. Lanes 1 and 2 correspond to pre- and post-cleavage respectively. The red arrow corresponds to the decrease in mass due to the removal of the His tag, and the blue arrow indicates a similar decrease in mass of what was thought to be an impurity.

protease, cleaved 6xHis tags and impurities. A low amount of elution buffer (corresponding to 20 mM imidazole) was then used to elute the MmyJ. TEV, tags and impurities were then eluted using 200 mM imidazole as shown in Figure 3.9a and 3.9b. Figure 3.9c explicitly shows the result of running another sample of His_6-MmyJ on an SDS-PAGE gel before and after cleavage. While it is obvious that the reduction in mass associated with the main bands between lanes 1 and 2 is due to the cleavage of the His tag from His_6-MmyJ, it was surprising to see that a similar decrease was also apparent in the fainter band at around 30 kDa. This was previously suspected to be an impurity such as SlyD, but the fact that it was cleaved indicates that this band corresponds to a dimer of His_6-MmyJ.

Whilst MmyJ is expected to form dimers, ArsR family dimers have always been reported to be hydrogen bonded [51], whereas for this dimer to have survived denaturing by the SDS within the gel it must be covalently bonded. What is interesting to note is that this band was thought to be an impurity as it was not always apparent; indeed it cannot be seen in Figure 3.4b. It was realised when quantifying different His_6-MmyJ preparations that the suspected dimeric band only appeared in high concentration samples. Hence, it was suspected that this covalent bond arose from the close proximity of His_6-MmyJ monomers when forced into concentrations higher than would be occur naturally, and thus could disrupt the biological function of MmyJ, causing difficulties in later experiments.

3.2 Site Directed Mutagenesis

3.2.1 Determination of Disulphide Bridge Formation

In order to investigate the suspected covalent bonding of His_6-MmyJ dimers, Dithiothreitol (DTT) was used as a reducing agent to disrupt any disulphide bridges that may be forming between monomers, as it was suspected that this would be the most logical reason for their formation. A 100 mM DTT solution was prepared and 0.3 µL was added to 100 µL 0.3 mM pure His_6-MmyJ, giving a 1:1 molar ratio. This was performed twice, with one mixture being left at room temperature and another being incubated at 90°C for an hour to encourage unfolding. Negative controls of His_6-MmyJwith dH_2O added in the same quantities as DTT solution were also prepared and incubated accordingly. After incubation, a 25 µL sample from each

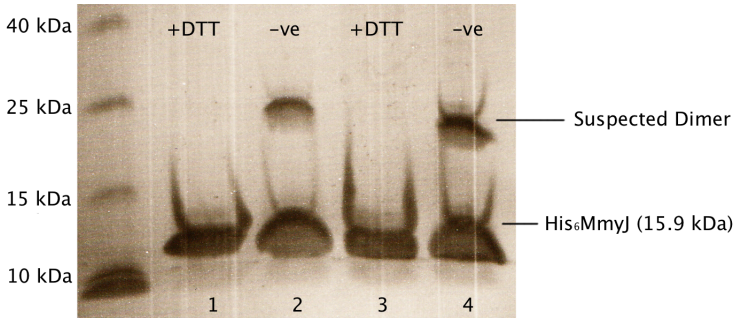


Figure 3.10: Native PAGE gel of His_6-MmyJ DTT assay investigating covalent dimer formation. Lanes 1 and 2 correspond to samples left at room temperature, and lanes 3 and 4 correspond to samples incubated at 90°C for an hour.

temperature was run against the negative controls on a Native-PAGE (i.e. non-denaturing) gel in order to see whether the addition of DTT had affected the suspected dimers. Figure 3.10 shows the result of this assay, from which it can be concluded that the addition of DTT clearly disrupts dimer formation as the bands associated with the dimer are completely removed. What is also apparent is that the dimer is not only robust to the denaturing properties of SDS (as seen in Figure 3.9), but is also resistant to thermal degradation. As it would be expected that the combination of the presence of SDS and incubation at temperatures this high would disrupt hydrogen bonded secondary structure, it can be concluded that this is a covalent dimer of His_6-MmyJ. It was thought that a disulphide bridge between monomers was the cause of the dimer formation, likely between two cysteine sites. After inspection of the amino acid sequence for His_6-MmyJ, it was found that the protein only contains one cysteine residue at position 49, and so it was proposed that mutating this residue could prevent disulphide bridge formation.

It is worth noting that, according to the previously reported homology model, this single cysteine lies at the start of the α3 helix, as shown in Figure 3.11. As such it could have a role in

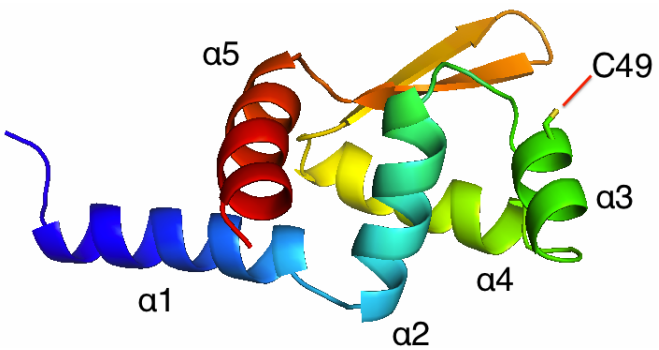


Figure 3.11: Previously shown homology model of MmyJ with the side chain of C49 shown explicitly on the outer edge at the start of the α3-helix.

stabilising the MmyJ:DNA complexes once formed, especially as it is on the outside of the helix in a position where it could conceivably interact with DNA. For this reason, it was decided that should the mutation be successful, work would continue with both the wild type and mutated MmyJ to ensure that functional investigations were not hindered by this mutation.

3.2.2 Mutation of C49

Site directed mutagenesis was performed to mutate the single cysteine at residue 49 into a serine, thus removing the possibility of disulphide bridge formation. This required the replacement of the TGC codon with AGC, hence only a single point mutation of t145a was needed. A single colony grew after transformation with the mutated plasmid, and a glycerol stock was created. A liquid culture was grown and the plasmid was extracted and sent for sequencing. The sequencing data can be seen in Figure 3.12, where it can be seen that the t145a mutation of the *mmyJ* gene required was successful.

3.2.3 Overproduction of His_6-MmyJ C49S and DTT Assay

BL21 star cells were transformed with the C49S plasmid and His_6-MmyJ C49S was then overproduced in LB media in the same manner as the wild type protein. This was then purified using a new HisTrap column to avoid cross contamination. The FPLC trace and associated SDS-PAGE gel can be seen in Figure 3.13. As this initial run was only to test the mutated plasmid, ensuring there were no other substitutions elsewhere that could disrupt expression of the mutated *mmyJ* gene, the overproduced protein was purified in a simple stepwise elution, leaving some impurities in the sample visible as faint bands in Figure 3.13b. It is apparent that the gene was indeed over expressed and the corresponding His tagged protein was overproduced but, as before, the band appears much lower on the gel than one would expect. Again this is suspected to be due to the shape of His_6-MmyJ, and further analysis by mass spectrometry can be seen in Section 3.3.

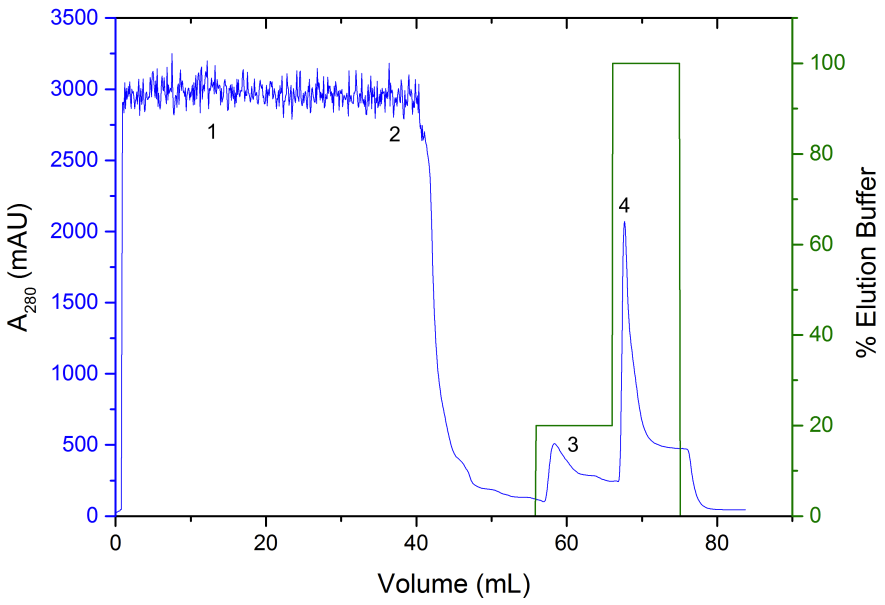
The same DTT assay was carried out with His_6-MmyJ C49S as with the previously expressed wild type His_6-MmyJ (previously shown in Figure 3.10). The results of this assay can be seen in Figure 3.14. While there are still some higher molecular weight bands apparent due to impurities previously mentioned, there is no apparent difference between the lanes with and without DTT

```
pET151mmyJC49S          ----------------------------------------------------
MmyJC49S-T7forwa      1 gtggagaattccccctctagatattttgtttactttaagaaggagatatac
mmyJ                    ----------------------------------------------------

pET151mmyJC49S        1 --atgcatcatcaccatcaccatggtaagcctatccctaaccctctcctc
MmyJC49S-T7forwa     51 atatgcatcatcaccatcaccatggtaagcctatccctaaccctctcctc
mmyJ                    ----------------------------------------------------

pET151mmyJC49S       49 ggtctcgattctacggaaaacctgtattttcagggaattgatcccttcac
MmyJC49S-T7forwa    101 ggtctcgattctacggaaaacctgtattttcagggaattgatcccttcac
mmyJ                    ----------------------------------------------------

pET151mmyJC49S       99 cgtggcggcacggatcacgacagagcgcatcaccgaccatccggacgctg
MmyJC49S-T7forwa    151 cgtggcggcacggatcacgacagagcgcatcaccgaccatccggacgctg
mmyJ                  1 -gtggcggcacggatcacgacagagcgcatcaccgaccatccggacgctg

pET151mmyJC49S      149 acgccatcaccctccagggcgtcctggacgcgctggtcgatccggtgcgc
MmyJC49S-T7forwa    201 acgccatcaccctccagggcgtcctggacgcgctggtcgatccggtgcgc
mmyJ                 50 acgccatcaccctccagggcgtcctggacgcgctggtcgatccggtgcgc

pET151mmyJC49S      199 cgcagcatcgtccggcagctggctaaggcacccgaggacatcgccagcgg
MmyJC49S-T7forwa    251 cgcagcatcgtccggcagctggctaaggcacccgaggacatcgccagcgg
mmyJ                100 cgcagcatcgtccggcagctggctaaggcacccgaggacatcgcctgcgg

pET151mmyJC49S      249 caccttcgacatcaccgtctcccgctcgaccggcactcaccacttcaagg
MmyJC49S-T7forwa    301 caccttcgacatcaccgtctcccgctcgaccggcactcaccacttcaagg
mmyJ                150 caccttcgacatcaccgtctcccgctcgaccggcactcaccacttcaagg

pET151mmyJC49S      299 tgttgcgccaggccgggatcatcaggcagtactacatcggcacctcgaag
MmyJC49S-T7forwa    351 tgttgcgccaggccgggatcatcaggcagtactacatcggcacctcgaag
mmyJ                200 tgttgcgccaggccgggatcatcaggcagtactacatcggcacctcgaag

pET151mmyJC49S      349 atgaacacgcttcgcaccgatgatctcgatcaggccttccccggcctgct
MmyJC49S-T7forwa    401 atgaacacgcttcgcaccgatgatctcgatcaggccttccccggcctgct
mmyJ                250 atgaacacgcttcgcaccgatgatctcgatcaggccttccccggcctgct

pET151mmyJC49S      399 caccgcgatcgtcgacgccgcggccagggagagctgaccggccaccgctc
MmyJC49S-T7forwa    451 caccgcgatcgtcgacgccgcggccagggagagctgaccggccaccgctc
mmyJ                300 caccgcgatcgtcgacgccgcggccagggagagctga------------

pET151mmyJC49S      449 gcccgcacggcgctccaaagggcgagctcagatccggctgctaacaaagc
MmyJC49S-T7forwa    501 gcccgcacggcgctccaaagggcgagctcagatccggctgctaacaaagc
mmyJ                    ----------------------------------------------------

pET151mmyJC49S      499 ccgaaaggaagctgagttggctgctgccaccgctgagcaataactagcat
MmyJC49S-T7forwa    551 ccgaaaggaagctgagttggctgctgccaccgctgagcaataactagcat
mmyJ                    ----------------------------------------------------

pET151mmyJC49S      549 aacccctttggggcctctaaacgggtcttgaggggtttttttgctgaaagga
MmyJC49S-T7forwa    601 aacccctttggggcctctaaacgggtcttgaggggtttttttgctgaaagga
mmyJ                    ----------------------------------------------------

pET151mmyJC49S      599 ggaactatatccggatatcccgcaagaggcccggcagtaccggcataacc
MmyJC49S-T7forwa    651 ggaactatatccggatatcccgcaagaggcccggcagtaccggcataacc
mmyJ                    ----------------------------------------------------
```

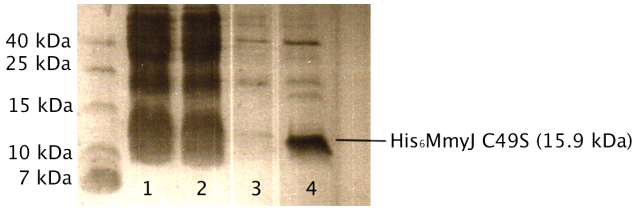
(Figure Continued Overleaf)

```
pET151mmyJC49S      649 aagcctatgcctacagcatccagggtgacggtgccgaggatgacgatgag
MmyJC49S-T7forwa    701 aagcctatgcctacagcatccagggtgacggtgccgaggatgacgatgag
mmyJ                    --------------------------------------------------

pET151mmyJC49S      699 cgcattgttagatttcatacacggtgcctgactgcgttagcaatttaact
MmyJC49S-T7forwa    751 cgcattgttagatttcatacacggtgcctgactgcgttagcaatttaact
mmyJ                    --------------------------------------------------

pET151mmyJC49S      749 gtgataaactaccgcattaaagctagcttatcgatgataagctgtcaaac
MmyJC49S-T7forwa    801 gtgataaactaccgcattaaagctagcttatcgatgataagctgtcaaac
mmyJ                    --------------------------------------------------

pET151mmyJC49S      799 atgagaattaattcttgaagacgaaagggcctcgtgatacgcctatttttt
MmyJC49S-T7forwa    851 atgagaattaattcttgaagacgaaagggcctcgtgatacgcctatttttt
mmyJ                    --------------------------------------------------

pET151mmyJC49S      849 ataggttaatgtcatgataataatggtttcttagacgtcaggtggcactt
MmyJC49S-T7forwa    901 ataggttaatgtcatgataataatggtttcttagacgtcaggtggcactt
mmyJ                    --------------------------------------------------

pET151mmyJC49S      899 ttcggggaaatgtgcgcggaacccctatttgtttatttttctaaatacat
MmyJC49S-T7forwa    951 ttcggggaaatgtgcgcggaacccctatttgtttatttttctaaatacat
mmyJ                    --------------------------------------------------

pET151mmyJC49S      949 tcaaatatgtatccgctcatgagacaataaccctgataaatgcttcaata
MmyJC49S-T7forwa   1001 tcaaatatgtatccgctcatgagacaataaccctgataaatgcttcaata
mmyJ                    --------------------------------------------------

pET151mmyJC49S      999 atattgaaaaaggaagagtatgagtattcaacatttccgtgtcgcccttа
MmyJC49S-T7forwa   1051 atattgaaaaaggaagagtatgagtattcaacatttccgtgtcgcccttа
mmyJ                    --------------------------------------------------

pET151mmyJC49S     1049 ttccctttttgcggcattttgccttcctgtttttgctcacccagaaacg
MmyJC49S-T7forwa   1101 ttccctttttgcggcattttgccttcctgtttttgctcacccagaaacg
mmyJ                    --------------------------------------------------

pET151mmyJC49S     1099 ctggtgaaagtaaaagatgctgaagatcagttgggtgcacgagtgggtta
MmyJC49S-T7forwa   1151 ctggtgaaagtaaaagatgctgaagatcagttgggtgcacgagtgggtta
mmyJ                    --------------------------------------------------

pET151mmyJC49S     1149 catcgaactggatctcaacagcggtaagatccttgagagttttcgccccg
MmyJC49S-T7forwa   1201 catcgaactggatctcaacagcggtaagatccttgagagttttcgccccg
mmyJ                    --------------------------------------------------

pET151mmyJC49S     1199 aagaacgttttccaatgatgagcacttttaaagttctgctatgtggcgcg
MmyJC49S-T7forwa   1251 aagaacgttttccaatgatgagcacttttaa------------------
mmyJ                    --------------------------------------------------
```

Figure 3.12: Sequenced plasmid containing single t145a mutation resulting in C49S mutant. 'pET151mmyJC49S' is the designed construct, 'MmyJC49S-T7forwa' is the result of sequencing the constructed plasmid extracted from transformed Top10 cells and 'mmyJ' is the sequence for the *mmyJ* gene, indicating the transformed base. Only the first 1248 bases of the pET151 plasmid are shown here.



(a)



(b)

Figure 3.13: (a) FPLC UV absorption trace at 280 nm of His_6-MmyJ C49S purification with corresponding % elution buffer, using wash/binding buffer with 10 mM imidazole and eluting in two steps. (b) 15% SDS-PAGE gel showing proteins present in collected fractions. Lane numbers correspond to volumes at positions indicated by numbers on UV trace.

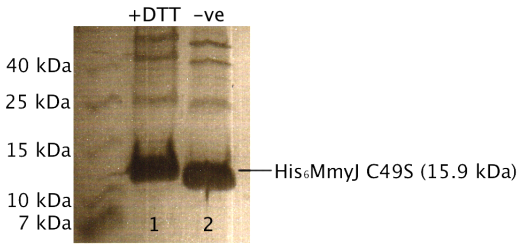


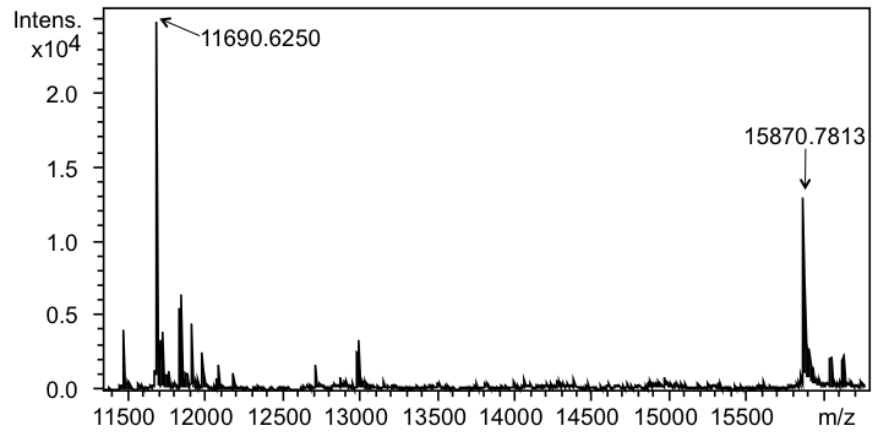
Figure 3.14: Native PAGE gel of His_6-MmyJ C49S DTT assay investigating covalent dimer formation. Samples were incubated at room temperature for an hour.

added. From this it can be concluded that there is no covalent dimer forming in this instance, and thus also that it was in fact disulphide bonds between C49 on separate monomers causing covalent dimers. Under the assumption that this covalent bond is not biologically relevant, work was carried out throughout with both the wild type protein and C49S mutation.

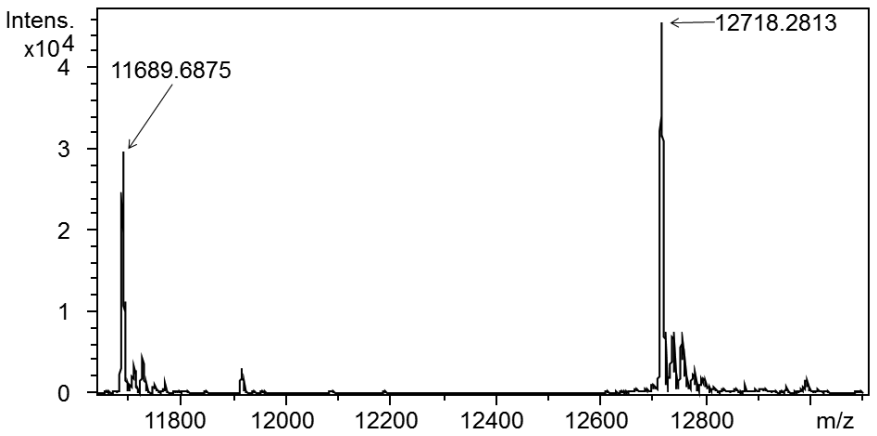
3.3 Mass Spectrometry Analyses

Mass Spectrometry (MS) was not routinely performed on each protein sample produced, however samples of His_6-MmyJ C49S and TEV cleaved MmyJ C49S were submitted in order to determine the molecular weight of the overproduced proteins. This was done to confirm that the overproduced MmyJ C49S matched the expected mass, thereby confirming that the peptide was intact. As the C49S mutant was created from the wild type expression vector, this would also confirm by inference that the wild type MmyJ was also overproduced correctly. In order to reduce the risk of uncontrolled fragmentation, and thus obtain a deconvoluted mass peak corresponding to the full length protein, electrospray ionisation was utilised [141]. The expected masses of His tagged and TEV cleaved MmyJ C49S are 15870.7573 and 12718.2474 Da respectively, calculated from the molecular formulae generated by ProtParam [102].

Figure 3.15 shows mass spectra of His_6-MmyJ C49S and TEV cleaved MmyJ C49S, where



(a)



(b)

Figure 3.15: Mass spectra of (a) His_6-MmyJ C49S and (b) TEV cleaved MmyJ C49S. Expected masses are 15870.7573 Da and 12718.2474 Da respectively.

it can be seen that the deviations between expected and measured molecular weight for the highest mass peaks are 0.0240 and 0.0339 respectively. These values can be converted to parts per million (ppm) as follows:

$$\text{Deviation } (ppm) = \frac{\text{Expected Mass} - \text{Observed Mass}}{\text{Expected Mass} \times 10^{-6}} \tag{3.2}$$

This gives deviations of 1.51 and 2.67 ppm for His tagged and TEV cleaved MmyJ respectively, well within the 5 ppm accuracy of the instrument.

From this it can be concluded that both variations of MmyJ overproduced by cells containing the engineered pET151 vectors do indeed contain the full length peptide chain. It is thought that the peaks with an m/z value of approximately 11690 in both spectra are either the result of fragmentation within the spectrometer or impurities within the instrument, as the purity of the submitted samples was confirmed by SDS-PAGE.

4 Stability Determination

As protein function is governed by its secondary and tertiary structure, it was desireable to ensure that the purified His_6-MmyJ was correctly folded, leading to further investigations into its stability.

4.1 Thermal Stability

In order to first check that the purified His_6-MmyJ was correctly folded and then investigate its thermal stability, circular dichroism spectroscopy was employed.

4.1.1 Circular Dichroism Spectroscopy

Circular Dichroism (CD) spectroscopy uses circularly polarised light to extract structural information about chiral molecules [142]. In the case of proteins, this can lead to quantification of the relative amount of α-helical, β-sheet or random coil character that comprise their secondary structure [143].

Circularly polarised light differs from linearly polarised (where the electric field oscillates only in one dimension perpendicular to the direction of travel of the light) in that the electric field vector rotates around the direction of propagation with the same frequency as the light; i.e. it takes one period of the base wavelength for the electric field vector to rotate through a full 360°. This vector may rotate either clockwise or anti-clockwise along the direction of propagation, defined as right-circularly polarised light (rpl) or left-circularly polarised light (lpl) respectively. As rpl and lpl are enantiomeric, they will lead to slightly different absorption behaviour when interacting with chiral molecules, and it is this difference in absorption A that gives a CD signal as follows [144]:

$$CD = \Delta A = A_{lpl} - A_{rpl} \tag{4.1}$$

NB: This shift in absorption is on the order of 10^{-4} times the absolute absorption, and so the light incident on the sample is modulated between rpl and lpl at the same frequency as the frequency of the light itself to enhance the signal received.

In proteins, the interactions between electronic transitions in adjacent peptide groups leads

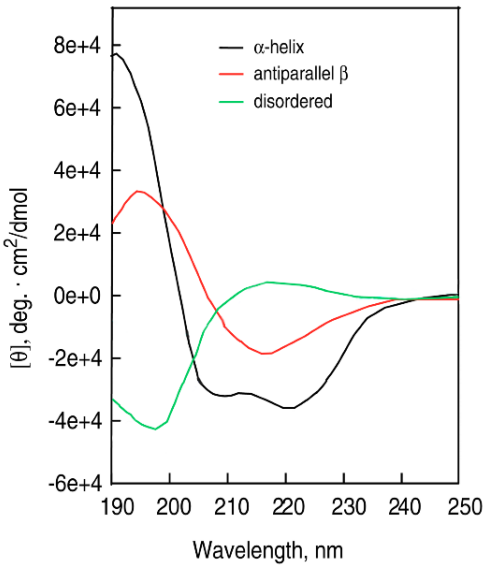


Figure 4.1: CD spectra of poly-L-lysine in α-helical (black), β-sheet (red) and random coil (green) conformations, demonstrating standard curves for those structures. Modified from [145].

to signature CD spectra depending on the secondary structure of the protein [144]. Figure 4.1 demonstrates the typical spectra one would expect from proteins that are completely α-helical (peak at 190 nm and double trough at 208 and 222 nm), β-sheet (peak at around 200 nm and trough at around 217 nm) or random coil (trough at around 200 nm). Most CD signals are measured in units of ellipticity θ for historical reasons, which can easily be converted to absorption using the relation $\Delta A = \theta/32.982$, with θ measured in degrees [142].

4.1.2 Protein Folding

Due to the high absorbance of NaCl in the high UV range [146], wild type His_6-MmyJ was prepared straight from purification by eluting with a variation on IMAC elution buffer containing $NaHPO_4$ substituted for NaCl. A sample of this was then taken and diluted to 0.5 mg/mL. A second sample was also prepared, this time of His_6-MmyJ C49S which had been dialysed into deionised water before being lyophilised and then resuspended in pH 6.5 phosphate buffer. The resulting CD spectra for both samples can be seen in Figure 4.2.

By comparison to Figure 4.1, it can be seen that Figure 4.2 demonstrates that both the wild type and C49S mutant of His_6-MmyJ have a high degree of α-helical structure as the spectra of both have the characteristic double trough at 208 and 222 nm as well as the peak at around 190 nm. This concurs with the predicted structure determined by homology modelling

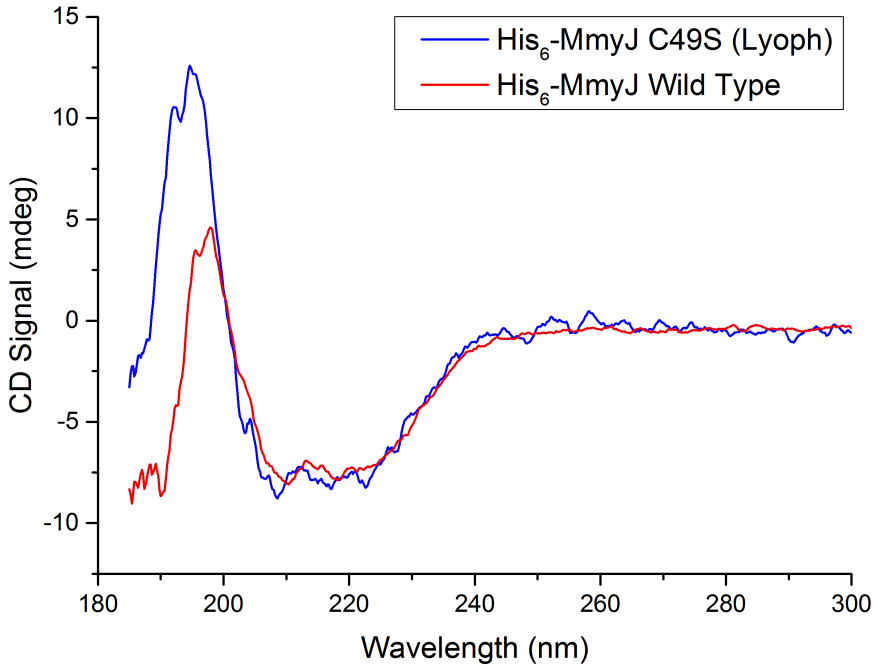


Figure 4.2: Circular Dichroism spectra of re-suspended lyophilised His_6-MmyJ C49S and native wild type His_6-MmyJ.

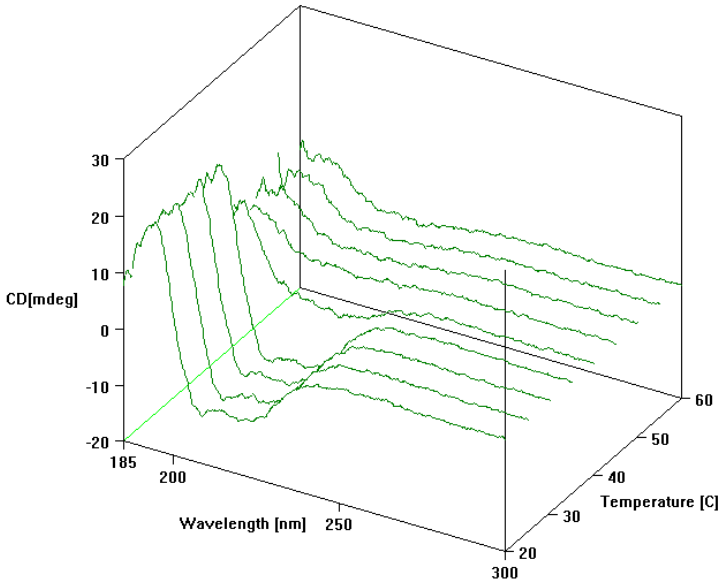
illustrated in Section 2.3, which classifies 65 of the total 111 amino acids as being in α-helices. This can be compared to only 14 amino acids being part of β-sheets which are likely to be responsible for causing the slight shift in α-helix peak closer to 200 nm.

Given the above analysis, it can be concluded that not only does recombinant His_6-MmyJ fold as expected, but also that the folding of His_6-MmyJ is not affected by the introduction of the C49S mutation. Also, it appears that the protein structure is robust to lyophilisation and rehydration after dialysis into deionised water, which means that this can be used as a sample preparation method in future if needed.

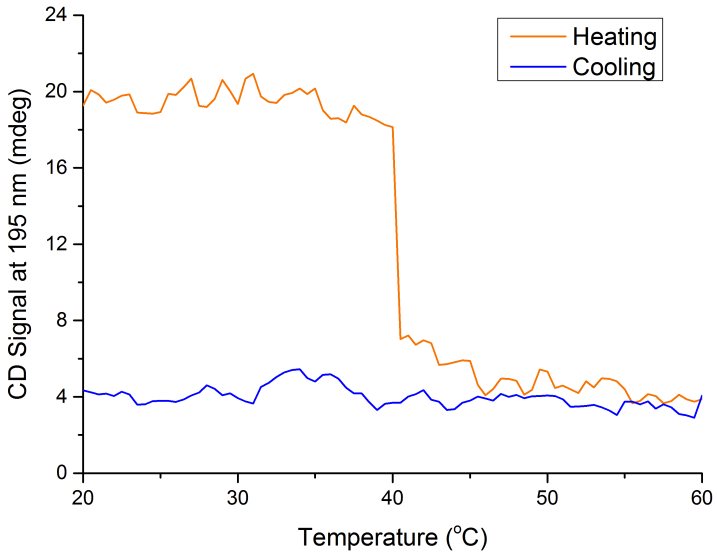
It is worth noting that the two spectra in Figure 4.2 were not taken on the same day and different cuvettes were used. This, along with the previously stated fact that they were in different buffers, is likely to have introduced minor changes in the profile of each sample, for example the relative height of the peak around 190 nm compared to the depth of the double trough, as well as a more pronounced shift in peak position for the wild type. It is not expected that these differences are a reflection in different folding between the wild type and mutant and, as this experiment was performed simply to see if the proteins were folded, it was not deemed necessary to repeat it.

4.1.3 Variable Temperature Assay

Once wild type and C49S His_6-MmyJ samples had been shown to be correctly folded, further work was undertaken to investigate the stability of the folded protein. The previously used sample of His_6-MmyJ C49S was re-used in order to investigate thermal stability. The instrument was set to raise the temperature by 1°C per minute from 20°C to 60°C inclusive, with the absorbance at 195 nm measured at each increment. As well as this, the instrument was set to record a full spectrum from 185 nm to 300 nm at 5°C intervals, requiring an equilibration



(a)



(b)

Figure 4.3: Data showing results of thermal stability assay performed on His_6-MmyJ C49S sample. (a) Shows how the entire CD spectrum varies at 5°C intervals, and (b) shows variation in CD signal at 195 nm when being heated and then cooled back down.

period of 5 minutes, as well as 5 minutes to record 3 acquisitions as before. The data from these experiments is shown in Figure 4.3.

Figure 4.3a revealed that the spectrum, and so the degree of folding, appears largely unchanged between 20°C and 35°C, with the features characteristic of α-helical folded proteins remaining the same. However, at 40°C it can be seen that the peak around 190 nm and the double trough at 208 and 222 nm all reduce drastically in amplitude, with these features disappearing from 45°C upwards, explicitly showing that His_6-MmyJ C49S denatures at around 40°C. This can also be seen in Figure 4.3b, where the absorption value at 195 nm also rapidly drops off at 40°C. It should be noted, however, that the absorption measurements, taken at the same temperature as the whole spectra (i.e. every 5°C), are taken at the end of the 5 minute equilibration period, and so the sharpness of the jump demonstrated in Figure 4.3b may have been exaggerated by the amount of time sat at 40°C. This does not alter the fact that the protein denatures at this temperature, but more implies that the line shape in Figure 4.3b would be different. Specifically, if this experiment was repeated without the equilibration periods or acquisition time for complete spectra, both of which were 5 minutes, the plot of absorption at 195 nm against temperature would likely have a less pronounced jump at 40°C. As such, it would probably follow the line of the curve that can be seen as the protein continues to denature from 41°C to 50°C. Also shown is a plot of data taken by repeating the measurements at 195 nm every minute as the sample was cooled back down to 20°C at the same rate, demonstrating that, as one would expect, the protein does not refold once denatured.

It is thought that as this method primarily monitors the stability of secondary structure, there would be no observable difference between the thermal stability of His_6-MmyJ C49S and the wild type protein. This is because the formation of a disulphide bridge between C49 residues in adjacent monomers is not thought to have any impact on the stability of α-helices and β-sheets elsewhere in the covalent dimer and, as such, the CD spectra observed here for the C49S mutant are expected to also reflect the behaviour of the wild type protein.

4.2 Protein Aggregation

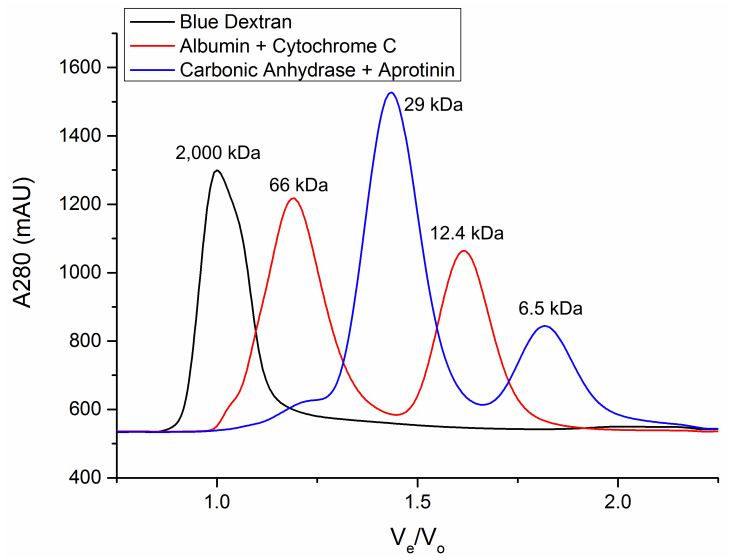
While attempting to identify the oligomeric state of His_6-MmyJ in its native form, it was noticed that the protein was aggregating, which is likely the reason many early experiments into the function and structure of His_6-MmyJ failed. Work was then performed to investigate this aggregation phenomenon.

4.2.1 Gel Filtration Chromatography & Calibration

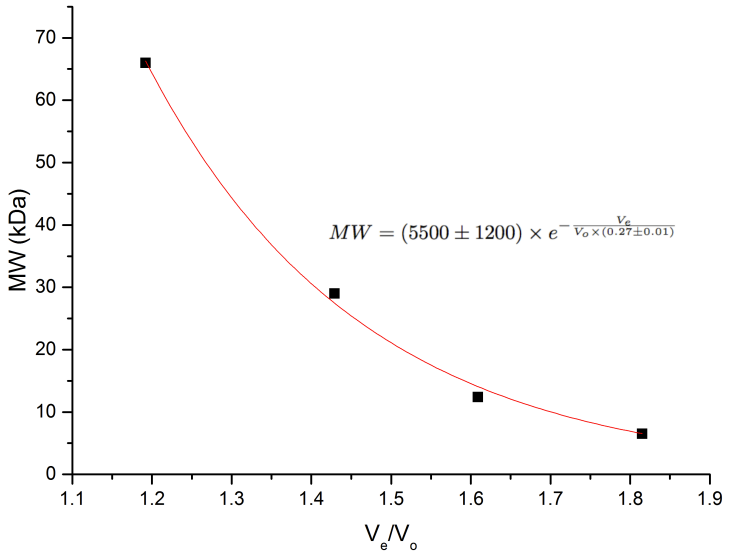
Gel filtration chromatography is a method for separating molecules in solution, based on their size, by passing the solution through a column containing a porous matrix, typically made of cross-linked beads [147]. Smaller molecules pass in between the beads whereas larger ones cannot enter the matrix fully and so elute from the column quicker, thus retention time on the column can be used to approximate the mass of molecules and fractions can be collected to separate the molecules in solution. The mass range and resolution varies from column to column depending on the polymers used to make the beads as well as bead and bed dimensions, pH and buffer contents. Also, there is some variation with protein shape, as a globular protein will pass through the column differently to an elongated protein of the same mass, but in this case it was assumed that His_6-MmyJ could be treated as globular, as supported by the previously described Phyre2 model.

Calibration data for the gel filtration column used to investigate oligomerisation of His_6-MmyJ can be seen in Figure 4.4. This was repeated for both FPLC machines available and was found to be near identical, so only the ÄKTApurifier 10 traces are shown. The peak for Blue Dextran corresponds to the so-called dead volume V_o, corresponding to the elution volume required for molecules too large to enter the matrix. Typically, data from gel filtration is shown with elution volume V_e normalised against V_o to reduce any dependance on the system used, as is done here.

It can clearly be seen from Figure 4.4a that the elution volume does increase with decreasing molecular mass as expected, and from Figure 4.4b it can be seen that this follows an exponential decay, in agreement with the predicted curve in the Technical Bulletin supplied with the calibration kit.



(a)



(b)

Figure 4.4: Calibration data for the Superdex™ 75 5/150 GL column on the ÄKTApurifier 10 system. (a) shows the UV chromatograms of the calibration solutions used while (b) shows how the molecular mass affects elution volume. The trend line was fitted using OriginPro 9.1 64Bit.

4.2.2 Oligomeric State

A 50 µL sample of His_6-MmyJ in IMAC wash binding buffer was run through the gel filtration column. It was decided that the initial run should be done in this buffer with glycerol still present, rather than the gel filtration buffer, for the sake of the stability of the protein. In order to determine whether this would have any effect on the peak positions, another run was performed with albumin and aprotinin added to the His_6-MmyJ sample to observe any shifts due to the inclusion of glycerol when compared to the calibration data. These data are shown

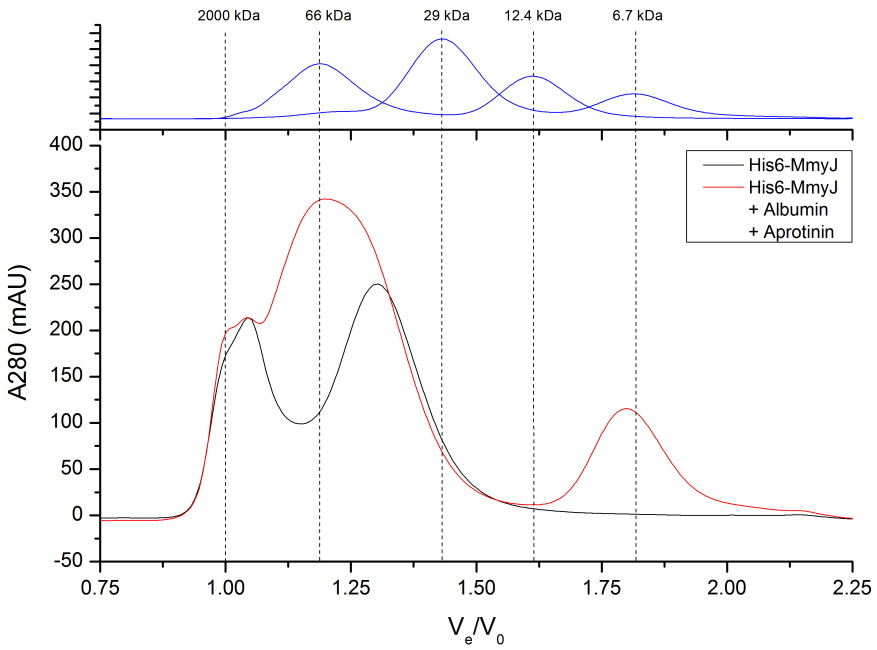


Figure 4.5: UV A280 chromatograms of 50 µL samples of His_6-MmyJ with and without addition of calibration proteins to observe peak shift due to presence of glycerol. Monomeric His_6-MmyJ has a molecular weight of 15.9 kDa, and so dimeric His_6-MmyJ is expected to have a peak corresponding to approximately 32 kDa.

alongside the original calibration data in Figure 4.5.

The first thing to notice about the His_6-MmyJ trace is that there is a large amount of material eluted in the dead volume peak. It was initially considered that the two peaks were the monomer and dimer of His_6-MmyJ, and that they had simply been shifted by the different buffer conditions. However, the second trace proves that this is not the case as while the albumin and aprotinin peaks are shifted slightly compared to the calibration peaks, the degree of difference is negligeable compared to the shift that would be needed for peaks of around 15 kDa and 30 kDa to appear where the two His_6-MmyJ peaks are. Also, the dead volume peak appears as a shoulder on the albumin peak at 66 kDa, demonstrating that it is not buffer conditions that have caused this peak to appear where it is. Due to this, the idea that the two peaks represented the monomer and dimer was discarded, and it was then assumed that the right hand peak corresponded to the dimer and the left hand peak corresponded to an unknown oligomeric state that was forming in solution from monomers and dimers, such that it was larger than 70 kDa and hence above the resolving range of the column[7]. In this case there

[7]Both native and SDS-PAGE gels were attempted with fractions collected from both peaks but, due to small initial sample size and the protein being diluted further upon elution, neither type of gel offered conclusive visualisation of the contents of these peaks.

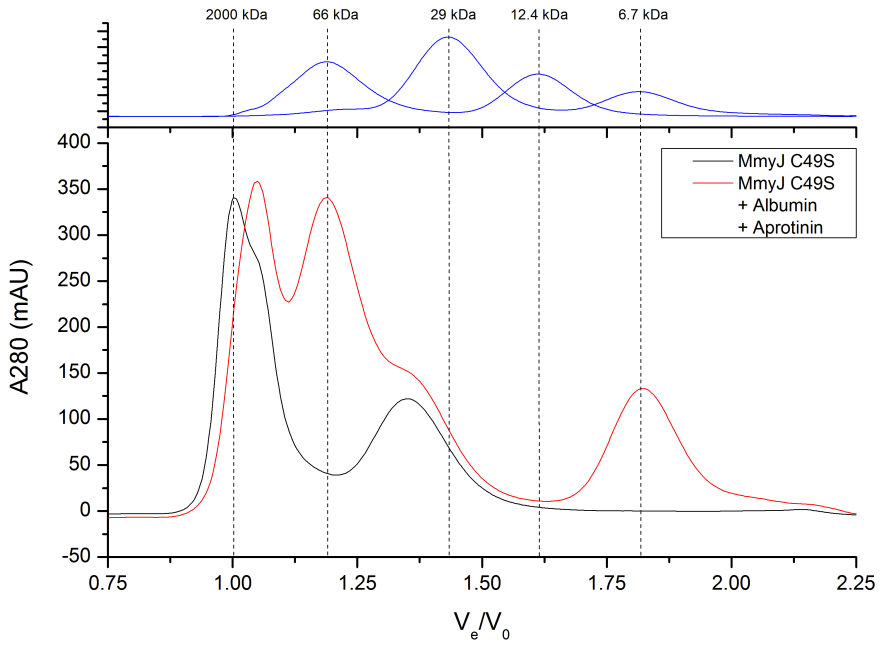


Figure 4.6: A280 chromatograms of 50 µL samples of His_6-MmyJ C49S with and without addition of calibration proteins, run under identical conditions as previous wild type sample.

was no observed peak assigned to the monomer, and so it was assumed that His_6-MmyJ does not appear in its monomeric form while in a non-denaturing solution.

The experiment was then repeated with a sample of His_6-MmyJ C49S prepared in the same manner, as shown in Figure 4.6. As can be seen, the His_6-MmyJ peaks appear to be in the same place, and so the mutation appears not to have altered the formation of oligomers. However, the dead volume peak is much larger, indicating that more protein has aggregated in this instance. This was not expected as the C49S mutation was introduced to minimise what was thought to be non-biologically relevant binding between monomers. Also, the dead volume peak is shifted significantly this time between the samples with and without added calibration proteins, yet the buffer conditions were identical for both runs, and so it can be inferred that this deviation may simply be due to a systematic error within the apparatus. If this is indeed the case then the minor aberrations apparent in Figure 4.5 may also be due to the same systematic error. From this and other runs where dead volume peaks were shifted slightly (a total of 22 separate experiments), the average shift in dead volume peak position was found to be ± 0.0255 when compared to the calibration value of $V_e/V_o = 1$. This error could be introduced into the system at a number of points; for example if a small air bubble was injected into the column along with the 50 µL sample, or if a slightly different length of tubing was used to attach the column

to the FPLC machine (although this was actively avoided by attempting to use the same piece of tubing for each experiment).

With this uncertainty in mind, the gel filtration data was then used to approximate the mass of the putative dimeric state corresponding to the right hand elution peak of both Figures 4.5 and 4.6. The position of this peak was recorded over 16 identical samples to give a mean peak position of 1.25 ± 0.04, where the uncertainty is the standard deviation across the 16 values. This is higher than the previously estimated systematic error, possibly due to there being more concordance between the elution time of large molecules that never enter the matrix, and so the actual deviation in elution time may in fact increase as molecule size decreases. Nevertheless, using this value and standard error propagation techniques [148] [149], along with the equation of the trend line from Figure 4.4b, the peaks were found to correspond to an approximate mass of 53 ± 11 kDa. This is not the expected mass of a His_6-MmyJ dimer, which should be 31.8 kDa. This has led to some speculation that His_6-MmyJ may form either a trimer or tetramer (as masses of 45 kDa and 60 kDa do fall within the error boundaries of this calculated value). However, despite there being some instances of tetrameric structures being reported [150], almost all ArsR family proteins are reported to form homodimers [51], and so more precise data would be needed to confidently describe the oligomeric state of His_6-MmyJ.

4.2.3 Analytical Ultracentrifugation

In a further attempt to identify the two oligomeric states of His_6-MmyJ apparent from gel filtration, a sample was prepared for Analytical Ultracentrifugation (AUC). This method sediments a sample in solution by spinning at high speeds, leading to centrifugal forces of around $250{,}000 \times g$ [151]. In this way, a gradient is formed throughout the sediment with larger molecules towards the outside edge of the rotor. UV absorbance can then be used to obtain a profile of the masses of the molecules present. Sedimentation velocity experiments measure sedimentation coefficients from which the mass can be calculated for a range of molecules present, whereas sedimentation equilibrium experiments accurately measure the molecular masses, but require only a single species to be present.

A sample of His_6-MmyJ C49S was prepared and submitted to the Birmingham Biophysical Characterisation Facility (BBCF) for analysis. As it was uncertain how many different oligomers

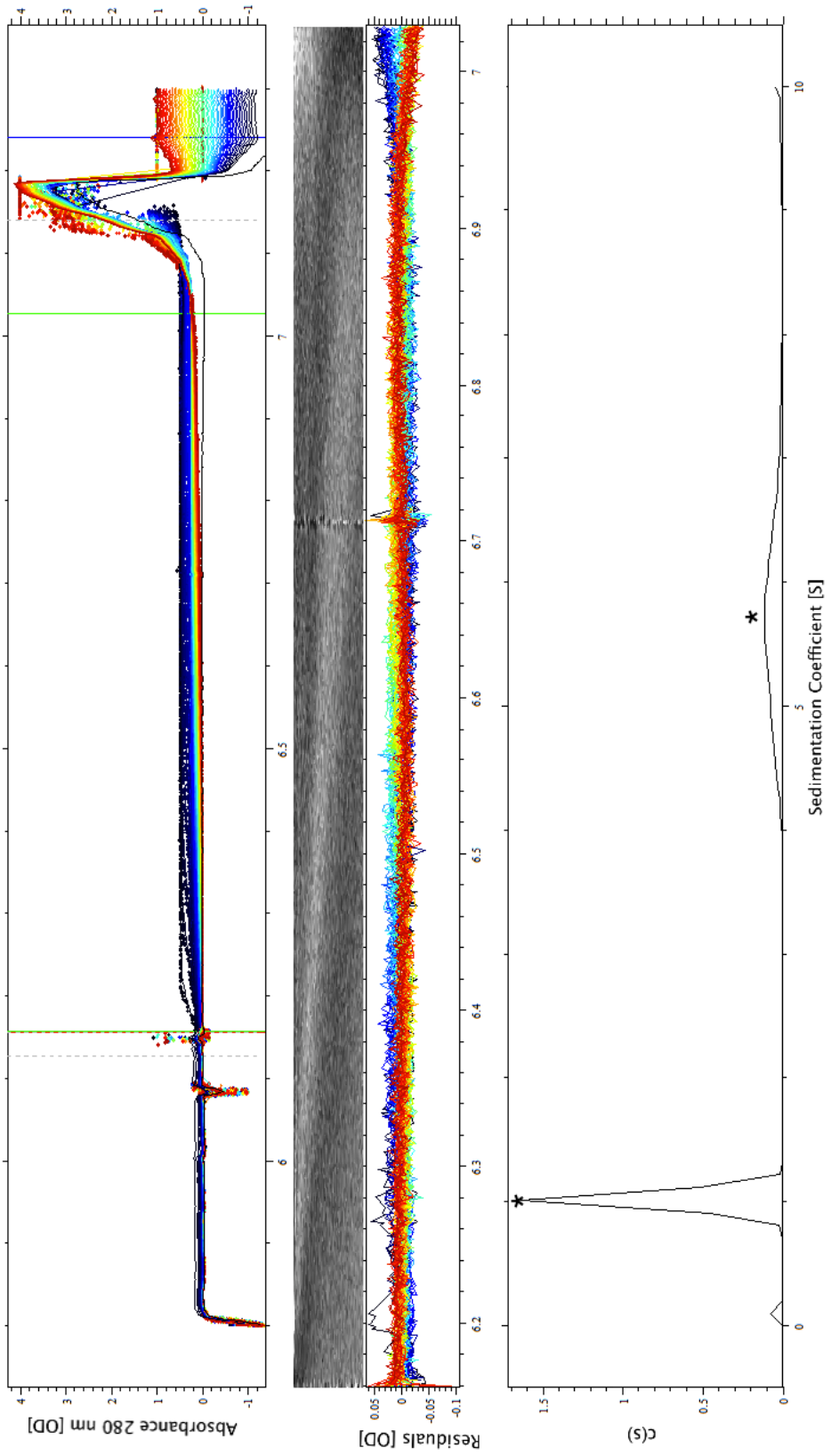


Figure 4.7: Data from AUC run of His_6-MmyJ C49S. The top panel shows the raw UV absorbance at 280 nm from 120 measurements, with the second panel showing the residuals from the fit of the data. The calculated distribution of sedimentation coefficient is then shown in the bottom panel, with major peaks identified by asterisks. The calculation was performed using values of $\rho = 1.0226$ (density), $\eta = 1.29984 \times 10^{-2}$ (viscosity) and $\bar{V} = 0.72024$ (partial specific volume) from Sedfit.

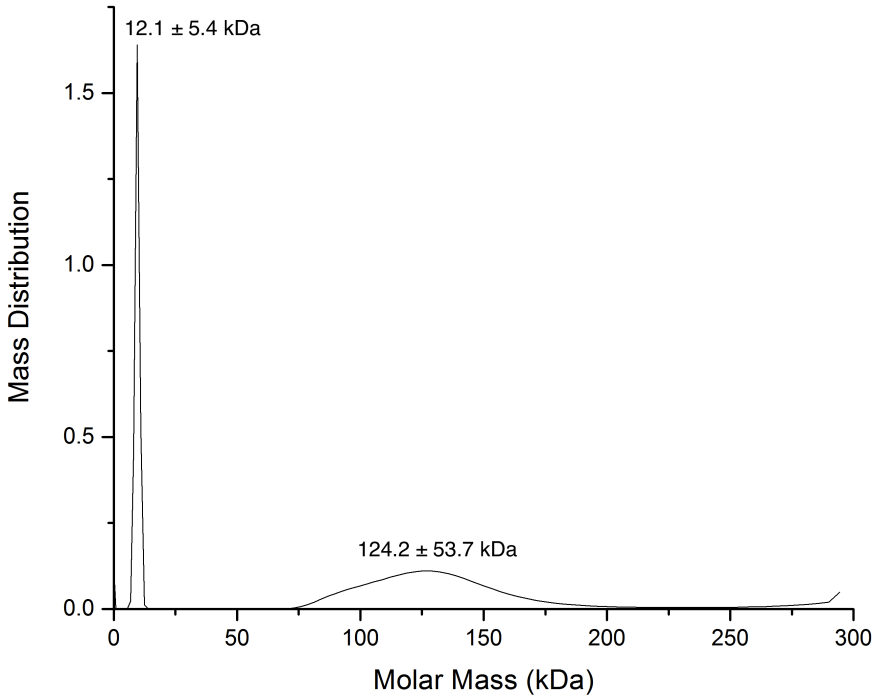


Figure 4.8: Mass distribution of species present in AUC sample, calculated from sedimentation coefficients according to instructions supplied with the data.

would be present, the sample was run in sedimentation velocity mode, sacrificing accuracy for breadth of results. The raw data from this, including absorbance spectra for multiple runs, quantification of residuals and distribution of sedimentation coefficients, can be seen in Figure 4.7. This analysis was done according to the instructions supplied with the data from the BBCF following a continuous distribution model [152]. It can be seen from the bottom panel of Figure 4.7 that there are two major peaks (annotated with an asterisk), corresponding to oligomers with different masses.

This distribution was then manipulated, again according to instructions supplied with the data, to transform the sedimentation coefficient values into masses. A plot of the resulting mass distribution can be seen in Figure 4.8. Using the full width at half maximum height as an approximation of the errors in peak position, the left hand peak corresponds to a mass of 12.1 ± 5.4 kDa while the right hand peak corresponds to a mass of 124.2 ± 53.7 kDa. As the monomer of His_6-MmyJ has a mass of 15.9 kDa, it can be concluded that the narrow peak on the left corresponds to the monomer, with the broad peak corresponding to the large aggregated state which cannot be resolved by gel filtration as it is too large to enter the matrix. The broadness of this peak also leads to the conclusion that there are several different oligomeric

states included in this aggregate. What is unexpected, however, is that there is no peak corresponding to a low order oligomer such as dimer, trimer or tetramer. This could be due to buffer conditions preventing the formation of hydrogen bonds between monomers, which then leads to the conclusion that the aggregate is likely formed by covalent bonds in order to still be present. With this in mind, more work was then carried out to investigate the formation of the aggregate using gel filtration.

4.2.4 Dynamic Equilibrium of Oligomeric States

In order to investigate the formation of the aggregated state, gel filtration experiments were carried out on collected fractions of His_6-MmyJ C49S. These fractions were then briefly centrifuged in spin filters to re-concentrate the eluted molecules, 50 µL of which were then taken and run through the gel filtration column a second time. Traces of these runs can be seen in Figure 4.9.

It is interesting to note that during several gel filtration experiments, including calibration runs, an extra peak was observed at high elution volume. As this must be a species much smaller than 6.6 kDa in weight by comparison to the calibration peaks, it was decided that it must be a small molecule, possibly an organic salt that had not fully dissolved, or free amino

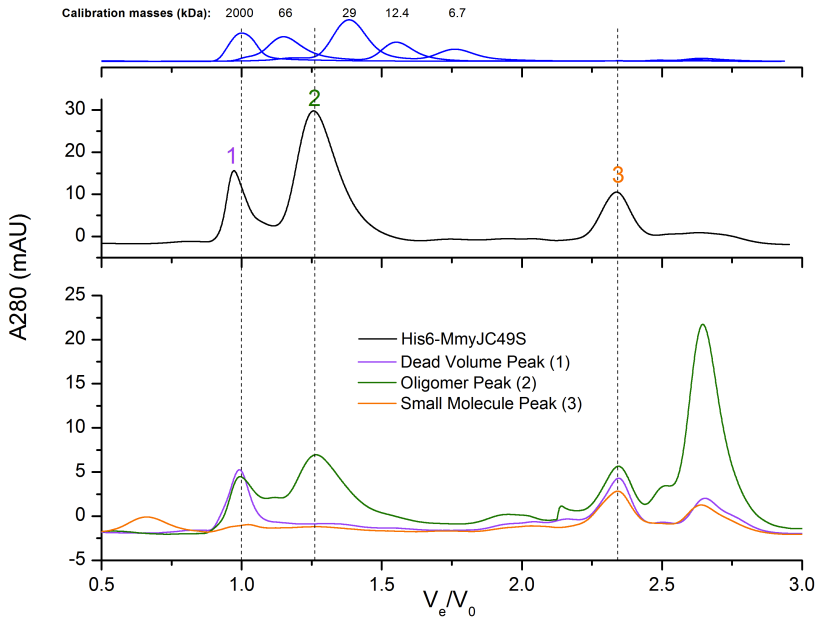


Figure 4.9: Gel filtration chromatograms of His_6-MmyJ C49S. The top panel shows the initial run, from which fractions were collected of peaks 1-3. These were then re-concentrated and run again, with these traces shown together in the bottom panel.

acids resulting from protein degradation. It is possible that some residual imidazole was still present in the sample, which does absorb at 280 nm and so could be responsible for the extra absorption peak, or that some residual plasticizers had entered the samples from plasticware used during preparation. For completeness sake it was included in the following analysis.

Figure 4.9 indicates that there is a dynamic equilibrium at play. The fraction collected from the dead volume peak does not then give peaks at any other positions when re-run, except the suspected small molecule, but the fraction from the oligomer peak of His_6-MmyJ C49S does give rise to a new dead volume peak, indicating that some of the oligomer had aggregated during the re-concentration process. Also, there appears to be severe broadening between the oligomer and dead-volume peaks, with a slight bump between the two, indicating another, possibly intermediate, phase present. This is thought to be evidence of the oligomers directly interacting with each other to form the higher order aggregates apparent in the AUC data. As suspected, the small molecule peak did not vary between runs, although another peak did become apparent at an even higher elution volume. However, these are not thought to be of interest.

At this point it was realised that the observed His_6-MmyJ aggregate could be formed by

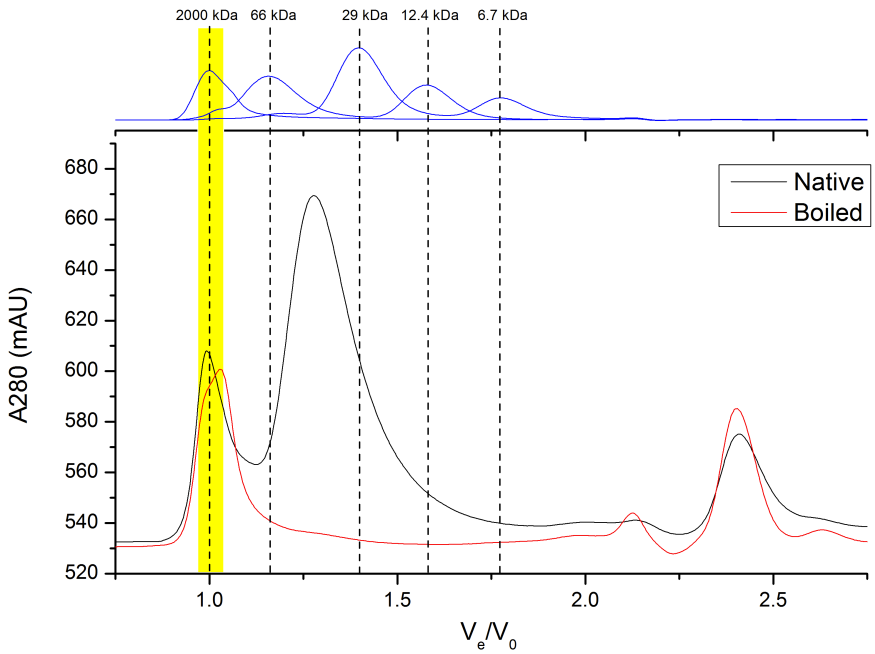


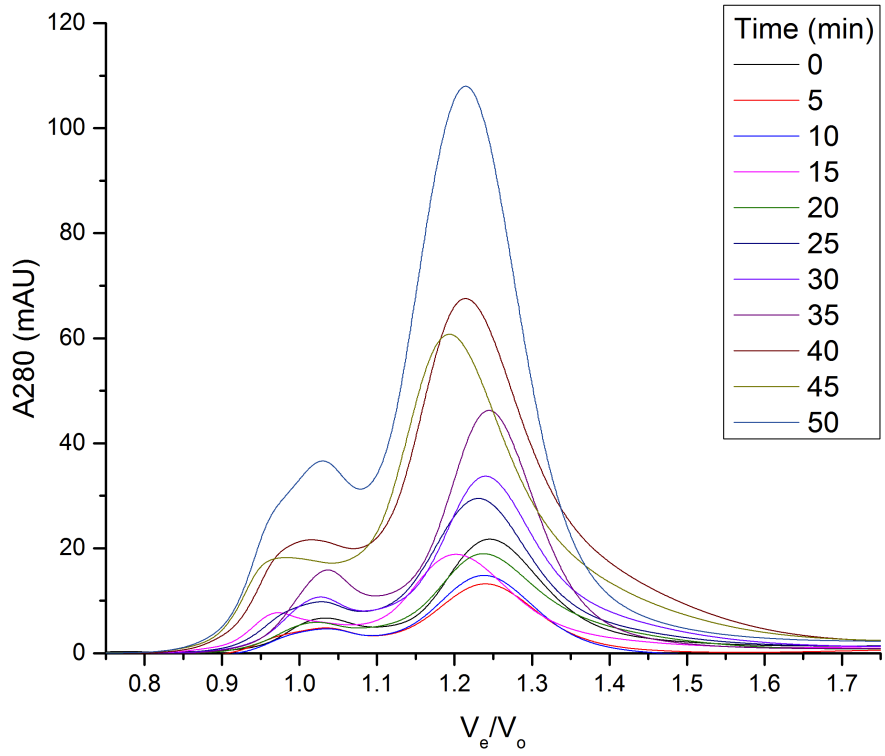
Figure 4.10: Gel filtration UV chromatograms of two runs taken from the same sample of His_6-MmyJ, before and after boiling at 95°C for 15 minutes. Region shaded yellow corresponds to approximate error in V_0.

the interaction of hydrophobic regions of unfolded portions of the protein [153], which would be a good explanation of why many preliminary structural and functional experiments had failed. In order to investigate this possibility, a sample of His_6-MmyJ was boiled at 95°C for 15 minutes, with 50 µL taken before and after boiling to compare the traces. These data can be seen in Figure 4.10. The absorption trace of the boiled protein matches the profile of the re-run dead volume peak in Figure 4.9, from which it can be concluded that the species responsible for the large dead volume peak observed in all gel filtration experiments so far is in fact from aggregated His_6-MmyJ after it has unfolded.

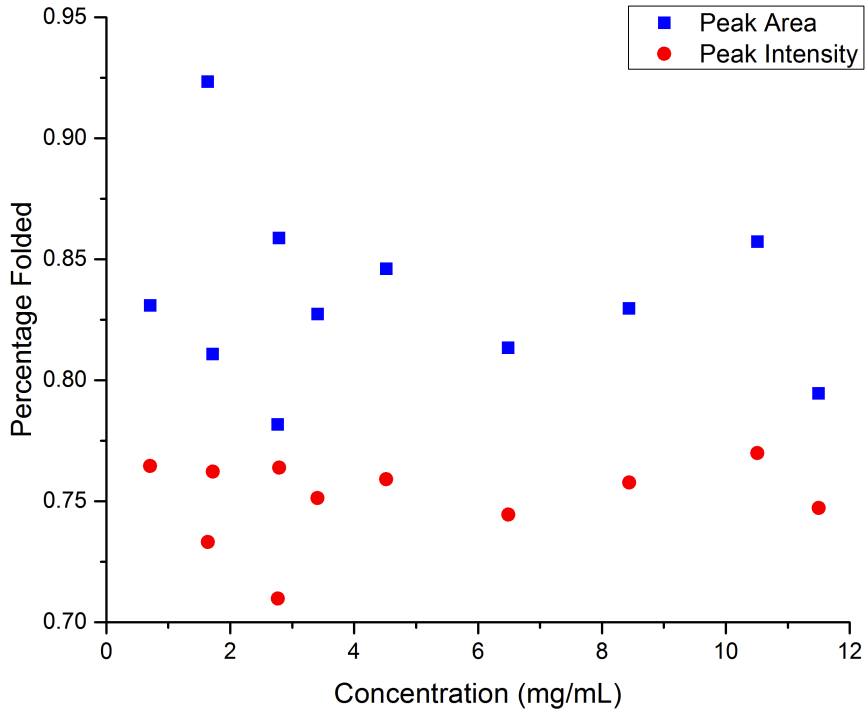
4.3 Concentration Stability

It was suspected that His_6-MmyJ might aggregate when concentrated to levels required for biophysical experiments, of the order of 10 mg/mL or approximately 600 µM - much higher than biologically relevant concentrations which are typically of the order of 1 nM [154]. In order to examine this possibility, an experiment was devised where the entire eluted amount of His_6-MmyJ from a 1 L culture, taken straight from the FPLC fraction collector after purification, was placed in a 10 kDa cut-off Amicon Ultra-15 Centrifugal Filter Unit and spun at 2,500×g at 4°C, with 200 µL taken at 5 minute intervals and kept on ice. The concentration of these fractions was then quantified on a Thermo Scientific NanoDrop Lite Spectrophotometer, before 50 µL of each was run through the gel filtration column previously used. In this way it was hoped that the ratio of intensity between the His_6-MmyJ oligomeric peak and the dead volume peak could be used to approximate the degree of unfolding, and hence determine whether concentration was a factor in this process.

The UV absorption traces for each fraction can be seen in Figure 4.11a. The general trend is that the peaks increase in both intensity and area as the protein gets more concentrated, as should be expected, but it is not immediately apparent whether the dead volume peak is increasing faster than the oligomeric peak. Figure 4.11b shows the estimated percentage folded for each concentration (established for each time period by taking a mean of 3 readings on the NanoDrop) by dividing both the area and maximum intensity of the oligomeric peak by the total of both the dead volume and oligomeric peaks together (assuming that all signals in



(a)



(b)

Figure 4.11: Results of investigating degree of unfolding as concentration is increased. (a) shows gel filtration traces at each time step, and (b) shows estimated folded percentage based on both peak area and intensity at different concentrations.

each sample were purely from His_6-MmyJ, whether it be in oligomeric or aggregated form). The percentage folded remains fairly constant to concentrations above 10 mg/mL, which is the highest concentration aimed for during biophysical work. Increasing concentration is therefore not causing the protein to unfold and aggregate, so the problem likely lies elsewhere in the preparation methods.

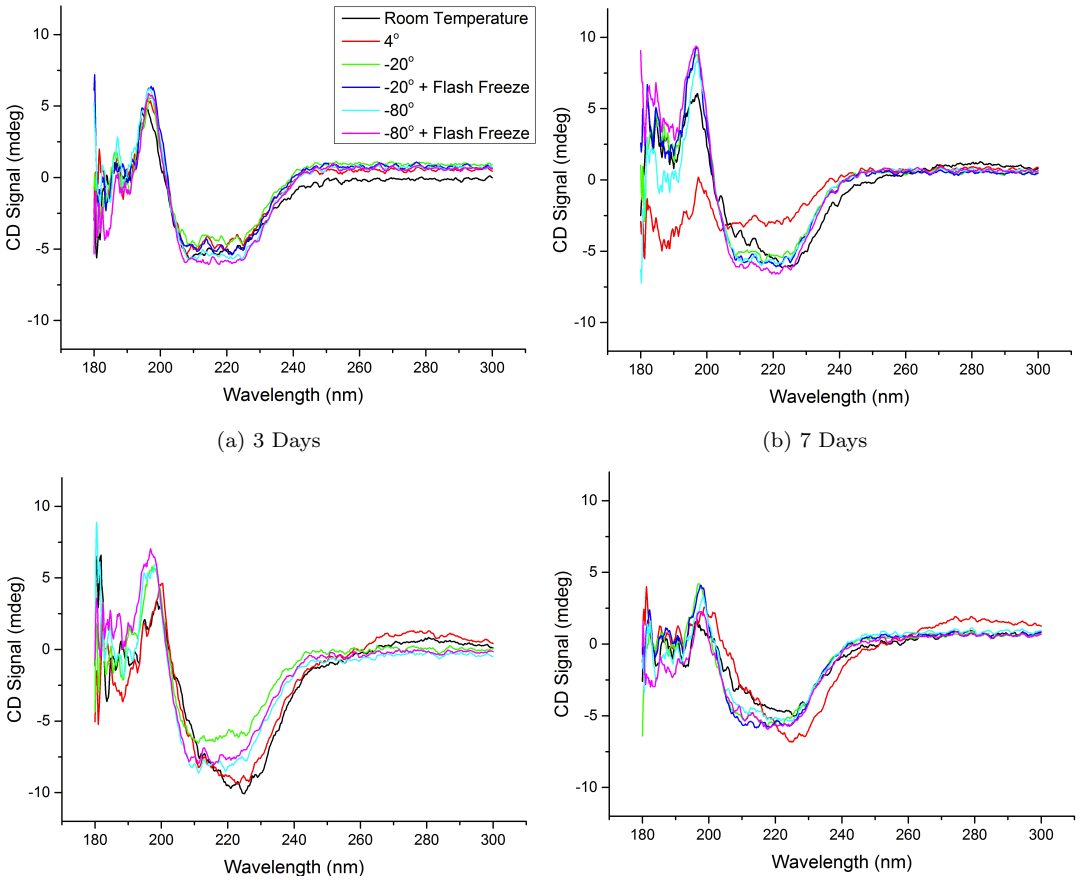
4.4 Cryo Stability

Stability under different freezing conditions was then investigated, with the aim of identifying the cause of the observed unfolding and aggregation.

4.4.1 Robustness to Prolonged Periods of Freezing

In order to investigate the effect of freezing His_6-MmyJ, 200 µL aliquots of freshly purified protein were kept under different conditions for different lengths of time before being thawed so CD analysis could be used as a gauge of how much the protein had unfolded. Figure 4.12 shows the resulting spectra for samples kept at room temperature, 4°C, −20°C (with and without flash freezing in liquid nitrogen) and −80°C (again, with and without flash freezing) for 3, 7, 19 and 28 days.

Figure 4.12 reveals that there is a degree of unfolding over time, demonstrated by a decrease in the amplitude of the characteristic α-helical peak at around 190 nm and double trough at 208 and 222 nm. However, the amount of unfolding is still a lot less than if the protein was boiled. What is also interesting to note is that, with the exception of the sample stored at 4°C for 7 days (which is thought to be an anomalous result, possibly due to poor cleaning of the cuvette used), there is very little difference in the amount of unfolding for the different storage conditions, even when comparing the protein that was simply left on the bench at room temperature for a month with that stored at −80°C. This is surprising as it implies that the protein is actually more stable than originally thought, and means that sample preparation need not necessarily be done on ice for future biophysical experiments. However, if this is the case then it still offers no explanation for why the protein was seen to be unfolding and aggregating in both the AUC and gel filtration experiments. It was considered that perhaps the protein was retaining some of its α-helical structure when mostly unfolded, leading to false



(a) 3 Days

(b) 7 Days

(c) 19 Days

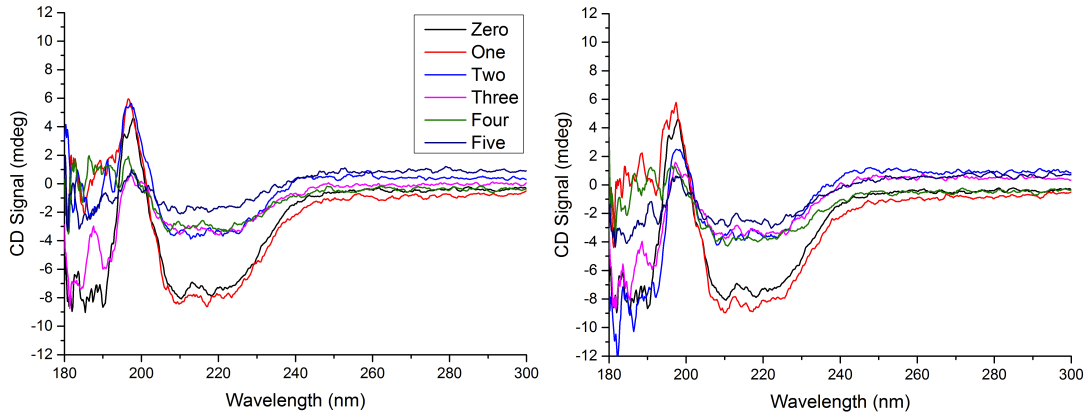
(d) 28 Days

Figure 4.12: CD spectra of samples stored at different temperatures, and with different freezing conditions for those stored below 0°C, taken after the samples had been stored for (a) 3, (b) 7, (c) 19 and (d) 29 days.

results in the CD spectra, but the previously demonstrated thermal stability assay clearly shows that when unfolded His_6-MmyJ loses all α-helical character. As such it can be concluded that the previously observed unfolding and aggregation is simply not happening in these samples, meaning there has to be another factor causing it.

4.4.2 Robustness to Repeated Freeze/Thaw Cycles

The next thing investigated was how the number of freeze/thaw cycles endured by the protein affected its structure. As before, CD spectra were taken of 200 µL aliquots of the same batch of protein after repeated freeze/thaws, as well as a baseline spectrum being taken before the samples were frozen. Also, for this experiment only samples stored at −80°C with and without flash freezing in liquid nitrogen were used. The results of this experiment can be seen in Figure 4.13, and what is immediately apparent is that while freezing His_6-MmyJ once has very



(a) (b)

Figure 4.13: Spectra of samples of His_6-MmyJ that have been frozen and thawed repeated times at (a) -80^oC only and (b) -80^oC with flash freezing in liquid nitrogen.

little effect, freezing it a second time after it has thawed seems to almost completely unfold the protein. This appears to be true whether the sample is frozen slowly or quickly in liquid nitrogen. From this it can be concluded that the issues experienced in attempted biophysical experiments were likely due to the protein experiencing multiple freeze/thaw cycles. This is in fact very likely as the -20^oC freezer that the His_6-MmyJ was generally stored in until low cryogenic stability was suspected has been noted to have problems closing due to a perished door seal leading to a build up of ice. Due to this, many aliquoted samples within that freezer will have inadvertently been frozen and thawed multiple times, hence causing disruption to experiments.

5 Function

Work began with purified recombinant MmyJ to investigate if it does indeed behave as a typical ArsR family transcriptional repressor.

5.1 DNA Binding

The first step towards proving ArsR-like functionality was to establish whether MmyJ binds to DNA and, if so, whether it binds specifically.

5.1.1 Proposed Binding Site

As previously mentioned in Sections 1.3.1 and 2.5, there is a region of DNA surrounding the promotor regions of both the *mmr* and *mmyJ* genes that has been shown to be protected by DNA fingerprinting [100], which was shown in Figure 2.14a. This region was found to contain a semi-conserved inverted dyad sequence, similar to the sequences to which ArsR proteins typically bind [28]. A comparison between this identified 13-1-13 dyad and the typical 12-2-12 motifs to which ArsR-like and SmtB-like family members are known to bind can be seen in Figure 5.1 [28]. Even allowing for the base shift due to the different sizes of the motifs, it can be seen that there is a greater alignment to the core ArsR-like motif than the SmtB-like motif. The two halves of the repeated sequence can clearly be seen to align in the case of the ArsR-like motif, whereas the SmtB-like motif is shifted to quite a large degree, with the centre points of the two motifs separated by a considerable number of bases. Incidentally, there are several other imperfect inverted dyad motifs within the same region, as highlighted in Figures 1.16b and 2.14a. However, these are two pairs of 7-3-7 dyads and a single 25-0-25 sequence, which

```
ArsR-like:      ---TAxAxTCAAAtax-xtaTTTGaxTxTA---
13-1-13 Dyad:   GACGGCTGTCAAA---C---TTTGATGGCCGTC

SmtB-like:      aAtAxxTGAacaxxtatTCAxaTxtt----------
13-1-13 Dyad:   --------GACGGCT-GTCAAACTTTGATGGCCGTC
```

Figure 5.1: Core DNA motifs recognised by ArsR-like and SmtB-like ArsR family proteins [28] (see Section 1.2.3 for details) aligned with imperfect dyad reported in [100]. For the core motifs, lower case letters indicate a lesser degree of conservation and 'x' indicates any base can be present at this position. Common bases are highlighted yellow. Bases in red indicate the non-repeated bases at the centre of the 12-2-12 and 13-1-13 sequences. Alignment performed using ClustalW2 [122].

align less favourably with the known core sequences than the 13-1-13 dyad, and so are not considered to be likely DNA targets for MmyJ.

Due to the similarity to the ArsR binding motif, the identified 13-1-13 repeat is considered the most likely target DNA sequence for MmyJ binding, and will be the focus of the following work.

5.1.2 Polymerase Chain Reaction Amplification of Intergenic Region

Polymerase Chain Reaction (PCR) amplification was used to amplify the 218 bp intergenic region between the *mmr* and *mmyJ* transcription start sites in its entirety from cosmid C73, containing the methylenomycin gene cluster from plasmid SCP1 [99]. Due to the shortness of the amplification region, Taq polymerase (error rate approximately 1 in 1000) was used with oligomer pair 4, listed in Section 8.1.3, which were designed using the Primer Design feature in Clone Manager 9 [155]. Later, the same PCR protocol was used to amplify 110 bp fragments of this intergenic region (bases 1-110 and 111-218) in order to locate the protein binding site, using primer pairs 5 and 6 (also designed using Clone Manager 9 [155]). The products of these amplifications can be seen in Figures 5.2a and 5.2b respectively. Both amplifications were successful, and all primers bound specifically, with no non-specific binding evident, which would lead to multiple undesired PCR products.

5.1.3 Electrophoretic Mobility Shift Assays

Electrophoretic Mobility Shift Assays (EMSAs), also called Gel Shift Assays, are a common method of demonstrating the specific binding of proteins to DNA [156, 157]. As with many

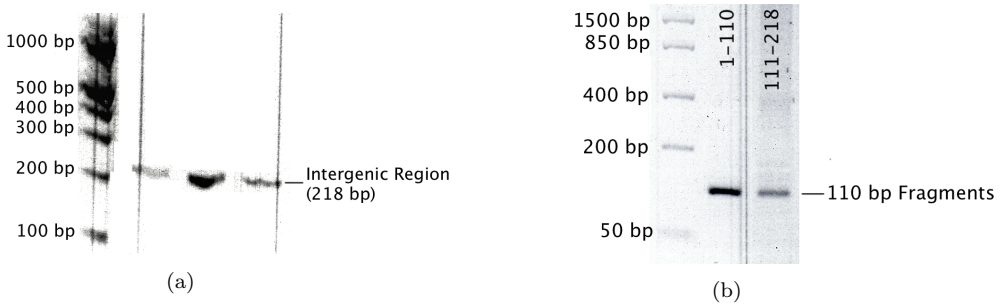


Figure 5.2: Products of PCR amplification of (a) entire 218 bp *mmr-mmyJ* intergenic region and (b) 110 bp fragments of the intergenic region (bases 1-110 and 111-218). Both sets of PCR products were run on 1.2 % agarose gels.

other gel electrophoresis techniques, the distance a compound moves through the gel when a voltage is applied varies with the charge and mass of the compound, with shape also having a contributing effect [158]. In EMSAs, there is typically a lane containing a specific sequence of DNA, as well as lanes containing the same DNA in a mixture with proteins that are suspected to bind to it. If a protein does indeed bind to the DNA sequence then, as the protein:DNA complex is larger and heavier than the unbound DNA, it would have reduced mobility and travel less distance through the gel in a given time. In this way, shifts in band positions can be observed between DNA and DNA/protein complexes if the protein binds to the DNA.

Likewise, this technique can be used when investigating ligands that cause the release of a bound protein from DNA. If a protein is known to bind to a DNA sequence, a lane containing this complex is run alongside lanes containing the complex plus ligands suspected of binding to the protein. If one of these ligands does in fact bind, then the DNA will be released and the band will be observed at the same position as the unbound DNA.

This process is illustrated in Figure 5.3, which simulates an experiment between two proteins binding to the same DNA sequence. A shift due to the addition of Protein 2 is observed and

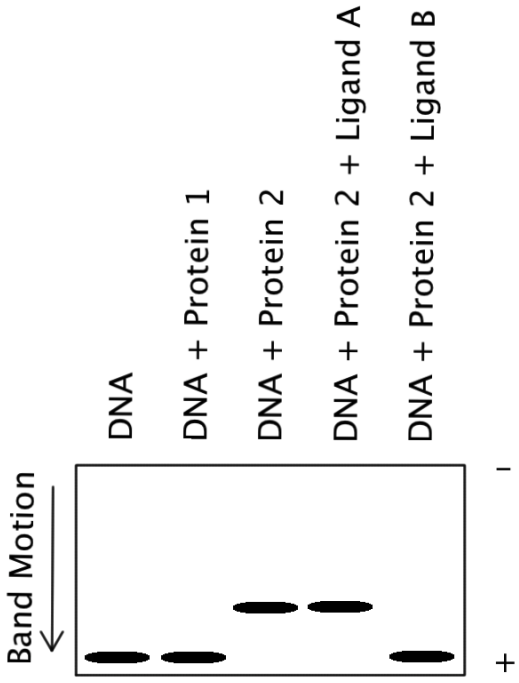


Figure 5.3: Simulation of EMSA in which the potential binding of two proteins to DNA is demonstrated, along with a simulation of the addition of binding ligands to the resulting protein/DNA complex. '+' and '-' indicate the potential applied across the gel.

reveals protein-DNA binding. When ligands A or B are added, it can be seen that ligand B prevents the band shift, and so implying that ligand B binds to protein 2 and causes it to release the DNA to which it was bound. This is the general principle of how EMSAs work, although there are of course some proteins that require a ligand to bind DNA, such as Fur, and so the procedure would differ in those instances.

5.1.4 Evidence of DNA Binding by MmyJ

Initial EMSA attempts were carried out using a 10 µL mixture of DNA and MmyJ (dissolved in IMAC wash/binding buffer) in an approximate 1:1 molar ratio. This was then added to 10 µL of binding buffer from Roche DIG Kit [159]. This mixture was incubated for 15 minutes before adding 5 µL non-denaturing loading dye, after which it was run on a 1% agarose gel containing a DNA intercalent. An attempted EMSA using the full 218 bp intergenic region is represented in Figure 5.4.

No binding is apparent with either the His tagged or cleaved MmyJ to the intergenic region as there is no shift upon addition of either protein. In order to improve the chances of successfully observing binding, further EMSAs were carried out on a 6% Native PAGE gel, using a recipe and protocol that had previously worked well for other users in the laboratory. Furthermore, upon investigation of the integenic region, it was noticed that it could be split in half; with one half containing the suspected binding site between bases 131 and 180 (incorporating the full possible binding site, not just the 13-1-13 inverted repeat) and the other proposed to be used

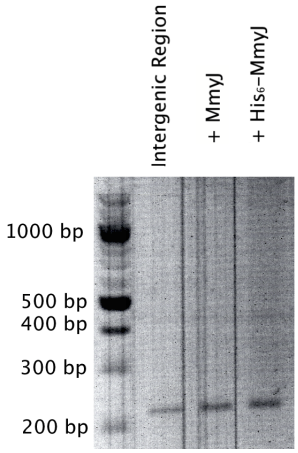


Figure 5.4: 1% Agarose EMSA investigating binding of both His tagged and cleaved MmyJ to complete 218 bp intergenic region.

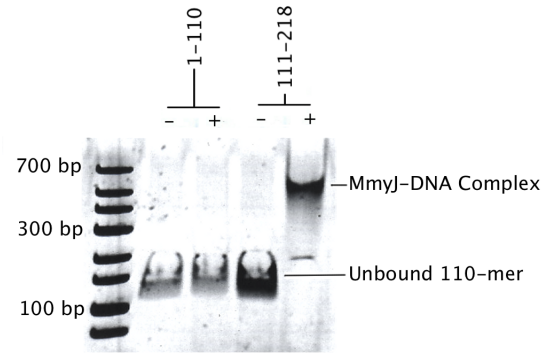


Figure 5.5: 6% Native PAGE EMSA showing binding of His_6-MmyJ to 110-mer corresponding to bases 111-218 of the intergenic region. - and + indicate absence and presence of MmyJ.

as an innate negative control, assuming MmyJ-DNA binding is specific. As such, primers were designed to divide the intergenic region, resulting in one 110-mer (bases 1-110) and one 108-mer (bases 111-218). These regions are numbered such that base 1 is the first base of the intergenic region, next to the *mmr* start codon. For simplicity, both will be classified as 110-mers for the remainder of this discussion.

Figure 5.5 shows the gel resulting from running the assay with the 110-mers. A clear shift is observed when His_6-MmyJ is added to fragment 111-218, but not fragment 1-110, indicating that His_6-MmyJ does indeed bind to DNA and that the binding is specific. As further proof of this, Figure 5.6 shows the SDS-PAGE gel from the purification of the sample of His_6-MmyJ used in this assay, from which the fractions corresponding to lane 4 were taken. The only bands present correspond to the pure His_6-MmyJ monomer and covalent dimer previously identified in Section 3.1.5 (NB: the dimer band is expected due to this being the wild-type MmyJ, and so the previously mentioned covalent binding is present). Hence, it can confidently be concluded that the binding observed in Figure 5.5 is due to His_6-MmyJ. It should be noted that although

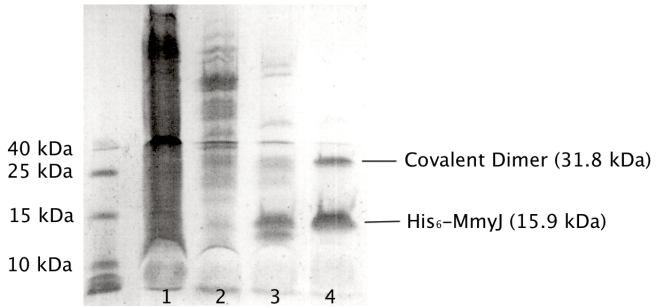


Figure 5.6: 15% SDS-PAGE gel showing purification of His_6-MmyJ sample used in EMSA shown in Figure 5.5. Lanes as follows: 1. Non-binding proteins, 2. Column wash, 3. First elution peak (His_6-MmyJ + impurities), 4. Second elution peak (pure His_6-MmyJ).

the His tagged form of MmyJ was used for this assay, there is no indication that such a tag alters the function of DNA binding proteins [160], and so it can be assumed that the DNA binding exhibited is in no way due to the presence of the His tag.

To further locate the protein binding site within the DNA fragment, pairs of 50 bp long oligonucleotides were designed and ordered from Sigma Aldrich. These oligonucleotides were designed such that the second half of the intergenic region was repeated in overlapping 50 bp blocks. In this way, there would be an additional overlapping oligonucleotide containing the entire binding site if it lay at the end of one oligonucleotide or bridged two adjacent blocks of 50 bases. As such, the region was divided into oligonucleotides covering bases 101-150, 131-180, 151-200 and 169-218, listed as oligonucleotide pairs 7-10 in Section 8.1.3. After mixing the pairs of complementary strands in a 1:1 molar ratio, these were then heated to 95°C and allowed to cool naturally back to room temperature, thus annealing the individual oligonucleotides into double stranded DNA.

Figure 5.7 shows the result of running an EMSA with the new 50-mers. A clear shift appeared in lanes containing both the 101-150 and 131-180 fragments, while weak binding is apparent in the lane containing the 151-200 fragment, and no binding in the lane containing 169-218. From this, it can be surmised that the protein binding site does indeed lie on the boundary between bases 101-150 and 151-200, and is therefore captured in its entirety by bases 131-180. This pattern of weak and strong binding can be overlaid with the predicted target sequence as shown in Figure 5.8; a modification of Figure 2.14a previously shown in Section 2.5.

With the DNA sequence annotated in this way, it can be seen that the 7-3-7 sequences,

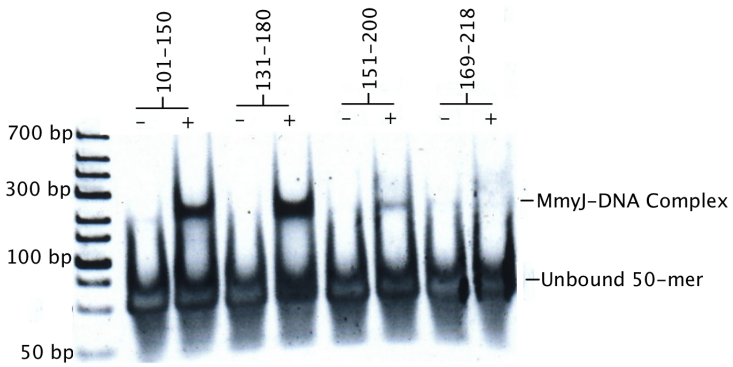


Figure 5.7: 6% Native PAGE EMSA showing binding of His_6-MmyJ to overlapping 50-mers covering the second half of the *mmr* to *mmyJ* intergenic region. - and + correspond to absence and presence of His_6-MmyJ.

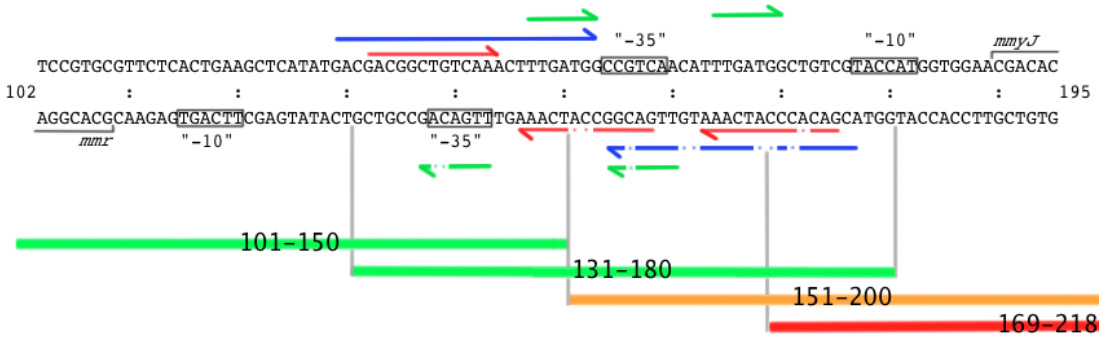


Figure 5.8: As Figure 2.14a: Fragment of intergenic region containing promoter sites for *mmr* and *mmyJ*, shown by black boxes. Colour coded arrows represent inverted repeats, with breaks in the arrows indicating imperfections in the repeated sequence. Bold lines added below the sequence correspond to 50-mers used in EMSA shown in Figure 5.7, with green indicating strong binding between the DNA fragment and His_6-MmyJ, orange indicating weak binding and red indicating that no binding was apparent.

indicated by green arrows, cannot be the protein binding sites as the 151-200 fragment actually contains one complete repeat and part of the other. Hence, if this was indeed the target sequence of MmyJ, then there should be strong binding exhibited by this fragment and weaker binding by the 101-150 fragment, which does not incorporate a single iteration of the full 7-3-7 sequence. Likewise, it cannot be the 25-0-25 sequence, shown by blue arrows, as if this was the case then the 131-180 fragment would show stronger binding than either 101-150 or 151-200, given that only the overlapping fragment contains the full 25-0-25 sequence, with the other two fragments simply contain half of the sequence. Thus, only the 13-1-13 sequence, labelled with red arrows, corresponds with the observed binding pattern, as 101-150 contains almost the full sequence and 131-180 contains the full inverted repeat. The fact that 151-200 binds weakly, despite containing the complete additional single copy of the repeated sequence as well as part of the true inverted sequence, can be explained simply by the observation that these two repeats are not antisense unlike the full inverted sequence between bases 132 and 158. As it has been shown that MmyJ binds specifically to DNA, if one of the repeats is the wrong way around then it does not match the specific requirements of the DNA binding site within MmyJ.

Bases	**Description**
101-129	No part of 13-1-13 sequence
117-145	First half of 13-1-13 dyad
131-159	Full 13-1-13 imperfect dyad
147-176	Second half of 13-1-13 dyad plus additional repeated 13 bp
161-189	Just additional repeat of 13 bp sequence

Table 5.1: Description of 29-mers with regards to the parts of the 13-1-13 repeat they contain.

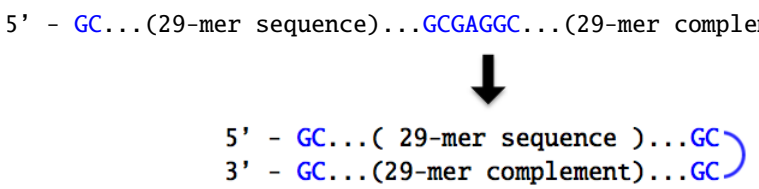


Figure 5.9: Design of self-annealing oligonucleotides [161]. Bases in blue form central loop and locking ends for 29-mer when cooled slowly from 95°C.

Also, the absence of binding to the 169-218 fragment indicates that MmyJ requires at least one complete half of the 13-1-13 repeat to bind.

Further work was then carried out using 29-mers, designed such that the binding of MmyJ to parts of the 13-1-13 sequence could be investigated. These are listed as oligonucleotides 11-15 in Section 8.1.3 and described in Table 5.1. These oligonucleotides contain additional bases at each end, as well as a loop section in their centre to promote self-annealing, with the 29 bp sequence and its complement placed on either side of the loop, as illustrated in Figure 5.9 [161]. These self-annealing oligonucleotides were heated to 95°C before being allowed to cool naturally to room temperature.

Figure 5.10 shows the corresponding EMSA using 1 µL of the self-annealed 29-mers diluted by a factor of 100. Also, in this instance, 18 µL of MmyJ was used to try and maximise interaction visibility. However, it can be seen that there is no clear binding, despite the expectation that there should be at least weak binding to the 117-145, 147-176 and 161-189 fragments, and strong binding to the 131-159 fragment. There is a slight shadow possibly indicating binding to the 131-159 fragment, but this is too faint to be conclusive and would represent a much larger shift than seen before. The lack of apparent binding could be explained,

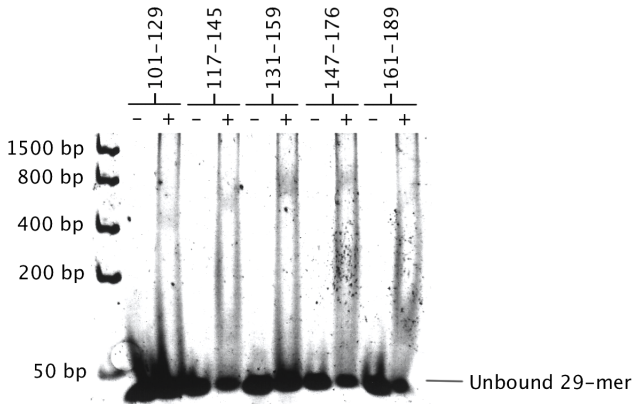


Figure 5.10: 6% Native PAGE EMSA using self-annealing 29-mers including parts of the 13-1-13 inverted dyad. - and + correspond to absence and presence of MmyJ respectively.

however, if MmyJ requires some of the bases surrounding the 13-1-13 repeat to stabilise the interaction with DNA. This could be the case if it proves to have a winged HTH structure or if the αR-αR' distance is shortened, requiring bending of the DNA to fully align with the major grooves as discussed in the case of HlyU in Section 1.2.5.

5.2 Ligand Sensing and DNA Release

Once DNA binding had been proven, the next step in the investigation into the function of MmyJ was to identify the binding ligand that would cause dissociation of the MmyJ:DNA complex, allowing the transcription of the *mmr* gene.

5.2.1 Evidence of Ligand Interactions with MmyJ

As with the interactions with DNA, EMSAs were used to investigate possible ligands sensed by MmyJ. Due to the absence of any genes indicating metalloregulatory function within the methylenomycin gene cluster, it was thought that either methylenomycin A (MmA) itself or one of its intermediates were likely ligands to which MmyJ would prove sensitive. Due to previously demonstrated binding of MmyJ to the *mmr* promoter region, it would make sense for the expression of *mmr* to be triggered by MmA or an intermediate being sensed by MmyJ, hence activating the self-resistance mechanism.

Figure 5.11 shows the result of an EMSA using a mixture of Methylenomycin C, D1 and

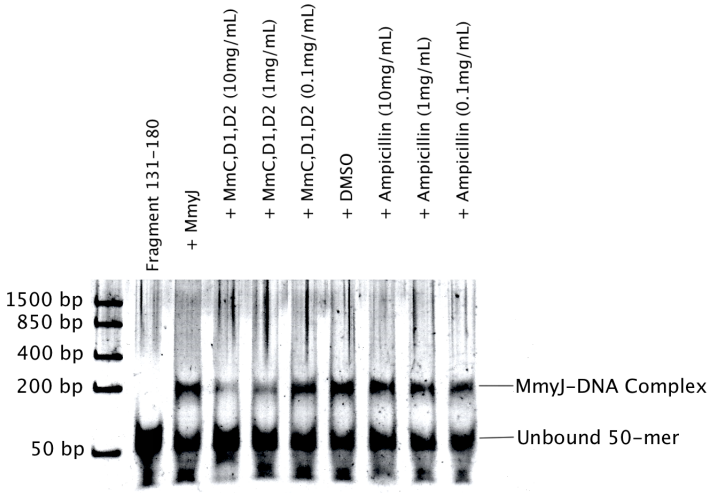


Figure 5.11: 6% Native PAGE EMSA showing sensitivity of MmyJ to mixture of MmC, D1 and D2 at different ligand concentrations, causing the DNA to be released. DMSO and ampicillin are used as negative controls.

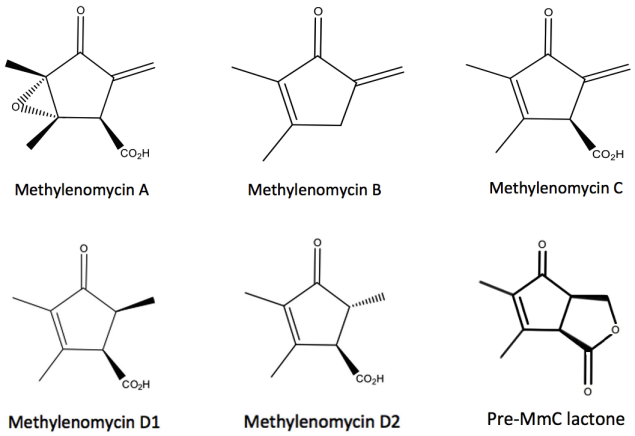


Figure 5.12: Structures of Methylenomycins A [89], B [163], C [90], D1 and D2 as well as the Pre-MmC Lactone [162].

D2 (where MmD1 and MmD2 are metabolic products of the saturation of the exomethylene side group in MmC [162]), obtained from Gideon Idowu, another member of the Corre research group. The structures of these intermediates are illustrated in Figure 5.12. As the MmC,D1,D2 mixture was suspended in Dimethyl Sulfoxide (DMSO), this is used as a negative control, along with ampicillin at the same concentrations as MmC,D1,D2, to ensure that any observed binding is specific to those compounds. It can be seen that while not sufficient to completely disrupt the MmyJ-DNA complexes, 10 mg/mL (approximately 50 µM) of the mixed MmC,D1,D2 does appear to reduce the amount of MmyJ bound to DNA, corresponding to approximately a 20 times molar excess. As this effect is reduced upon diluting the MmC,D1,D2 mixture to 1 mg/mL, and is not visible when diluted to 0.1 mg/mL, it can be concluded that one or more of the compounds MmC, MmD1 and MmD2 is sensed by MmyJ, disrupting the MmyJ:DNA complex. This is reinforced by the observation that neither DMSO nor ampicillin appear to inhibit the MmyJ-DNA interaction in the same manner.

Further samples of pure MmA, MmC and Pre-MmC Lactone were likewise dissolved in DMSO to a concentration of 10 mg/mL (50 µM) and tested. Figure 5.13 shows an EMSA performed using 1 µL of each of these compounds, along with the previous mixed sample of MmC,D1,D2. In this instance, streptomycin was used as a negative control, being an antibiotic also produced by *Streptomyces* species [164]. It can be seen that the reduction in DNA binding of MmyJ in the presence of pure MmC is comparable to the MmC,D1,D2 mix, implying that if more than one compound in this mixture is indeed sensed by MmyJ, MmC is probably

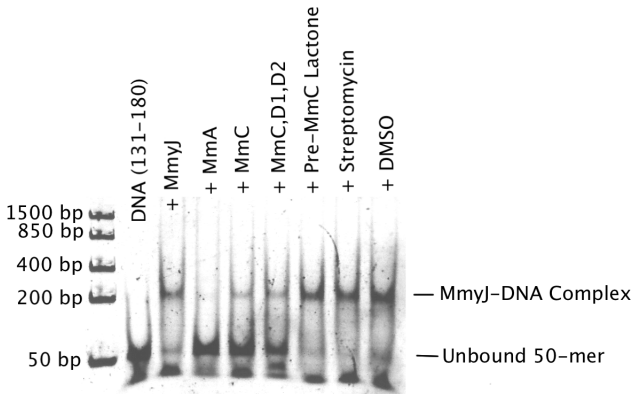


Figure 5.13: 6% Native PAGE EMSA showing sensing of MmA and MmC by MmyJ, causing the release of DNA. All ligands were added at a concentration of 10 mg/mL, with DMSO and streptomycin used as negative controls.

the dominant ligand. Interestingly, the addition of MmA almost completely disrupts the MmyJ:DNA complex, proving that MmyJ is more sensitive to this compound, and hence it is likely to be the most specific ligand of MmyJ. No sensitivity is exhibited to the Pre-MmC lactone, nor streptomycin, indicating that the interaction observed is specific to MmA and, to a lesser extent, MmC (which have a similar structure, as seen in Figure 5.12).

As a further study, the methylenomycin furan signalling molecules (MmFs) were also investigated as potential ligands to MmyJ. These are naturally produced by *S. coelicolor* A3(2) and are known to induce methylenomycin production by association with the TetR transcriptional repressor MmfR [99, 166]; their structures are shown in Figure 5.14. As they trigger the biosynthesis of methylenomycin, it is logical to think that they may also trigger its resistance mechanism, hence their inclusion in this investigation. Figure 5.15 shows an EMSA investigating

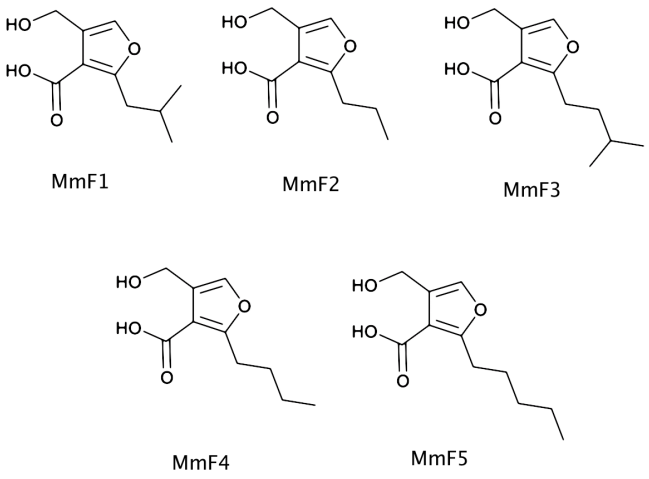


Figure 5.14: Structures of the 5 methylenomycin furan (MmF) signalling molecules [165].

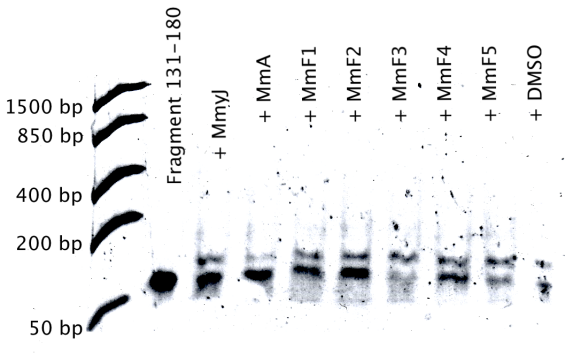


Figure 5.15: 6% Native PAGE EMSA investigating sensing of MmF compounds by MmyJ.

MmFs, which were obtained from the Challis Research Group and dissolved in DMSO to 10 mg/mL as before, with MmA used as a positive control. An apparent reduction in intensity of the shifted band upon the addition of MmA could be observed, but not for the MmF compounds, so they do not appear to be sensed by MmyJ. However, the results of this EMSA are slightly inconsistent with previous assays, as upon closer inspection it can be seen that the DNA binding of MmyJ in all lanes causes a drastically reduced shift when compared to Figures 5.7 and 5.13, despite the same conditions being used. There is still some interaction occurring, as there is a visible shift between the lane containing just DNA and the ones with MmyJ added, however the interaction appears to involve a smaller shift and, therefore, potentially a lower mass of protein. This could be the result of monomeric binding or partial proteolysis of MmyJ prior to purification, or indeed it could be the opposite case that this EMSA demonstrates true dimeric binding and previous results indicate a super-shift caused by the possible formation of a tetramer (as SmtB is known to do [67]). Multiple repeats of this assay have since been attempted using freshly expressed and purified protein, as well as freshly extracted MmA, with no improvement in clarity. As the observed MmA sensing in this case is not as efficient as previous assays, despite being used in the same molar ratio with MmyJ, the results of this experiment would need validating by further experimentation.

The same issue is also apparent in assays carried out at the same time investigating the sensitivity of MmyJ to metal ions known to bind to ArsR family proteins, as can be seen in Figure 5.16, which is also inconclusive. In this instance, the metal salts listed in Table 5.2 were dissolved in dH_2O to a concentration of 10 mg/mL. However, the shift of the MmyJ:DNA complex is again reduced and the positive control, (this time a mixture of MmA and MmC,

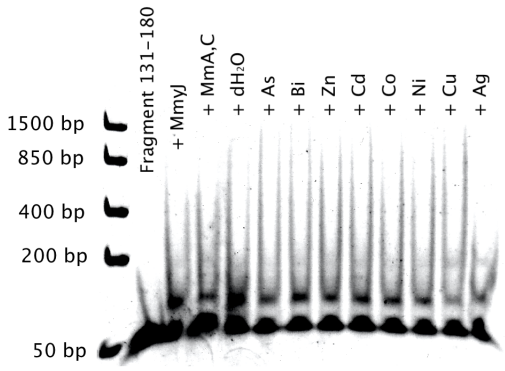


Figure 5.16: 6% Native PAGE EMSA investigating sensitivity of MmyJ to metal ions known to bind to other ArsR family proteins.

both at 10 mg/mL concentration), fails to be sensed by the MmyJ. Interestingly, in this EMSA there is a very faint band apparent in some lanes, most noticeably those with copper and silver added, at the same shift as previous positive results. It could therefore be interpreted that trace amounts of these metals aid in dimer formation. However, this assay used approximately a 20 times molar excess of metal ions to MmyJ, and so if this was truly the case then a large amount of metal ions would need to be present to aid MmyJ dimer formation *in vivo*.

This split between the previous shift and the reduced shift can be seen clearly in Figure 5.17, which shows an attempted EMSA investigating the effect of reducing the concentration of MmA added to the reaction mixture. A shift close to the 200 bp band on the ladder could be observed, as in previous figures, in addition to the shift lower down the lanes. It is interesting to note that the top band is removed to a greater extent by the addition of MmA, and only conclusively at a concentration of 10 mg/mL. This indicates that whatever MmyJ species is binding to the DNA and causing the band labelled as 'Half Shift' to appear only weakly senses MmA, explaining the comparative failure of the positive controls in Figures 5.15 and 5.16. The degree of shift between the unbound band and the Half Shift band seems to be approximately the same as

Metal	**Salt Used**
Arsenic	As_2O_3
Bismuth	$Bi(NO_3)_3.5H_2O$
Zinc	$ZnCl_2$
Cadmium	$CdSO_4$
Cobalt	$Co(NO_3)_2.6H_2O$
Nickel	$NiSO_4.6H_2O$
Copper	$CuSO_4.5H_2O$
Silver	$AgNO_3$

Table 5.2: Salts used as source of metal ions used in EMSA shown in Figure 5.16.

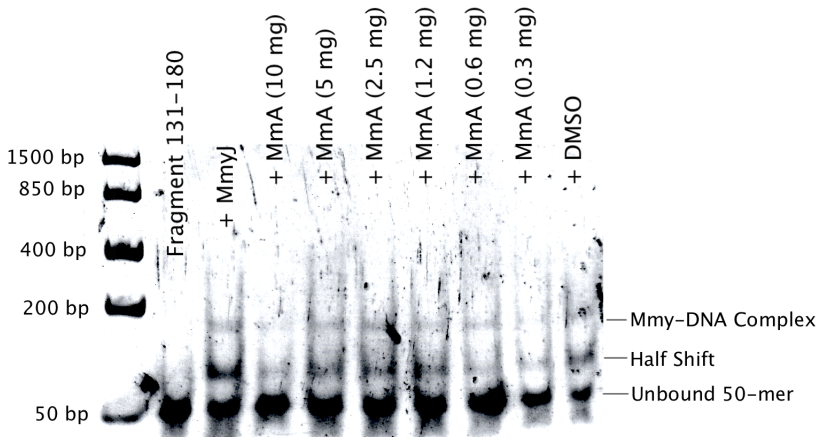


Figure 5.17: 6% Native PAGE EMSA showing the increase in preserved MmyJ:DNA complex as MmA concentration is reduced. Two distinct shifts are apparent; only one of which appears to be removed effectively by the addition of MmA. Note: only 5 µL of protein was added to each lane in this assay, so as to increase the effect of lower concentrations of MmA.

the degree between the Half Shift band and MmyJ-DNA complex band, with a slight deviation due to the non-linear nature of gel electrophoresis. This implies the previous suggestion, that this could be the result of either monomer/dimer or dimer/tetamer binding, may have some truth to it. Of these scenarios, the monomer/dimer equilibrium is perhaps most likely, as only one ArsR family protein has thus far been reported to exist in a tetrameric state, requiring two pairs of 12-2-12 dyads for DNA binding [67], where this system only has one full 13-1-13 dyad. If this is the case, then the lack of MmA sensing by what would be the monomeric band implies that the site sensitive to MmA lies in or around the dimeric interface of MmyJ, requiring both monomers for maximum sensitivity. At present, however, this is just speculation.

5.3 Functional Analysis *in vivo*

In an attempt to complement the *in vitro* results reported above, Reverse Transcription PCR was utilised and two distinct *in vivo* systems were also designed.

5.3.1 Reverse Transcription PCR

Reverse Transcription PCR (RT-PCR) is a method that can be used for quantifying the transcription of mRNA in cells, and hence determine whether certain genes are expressed [167, 168]. This procedure involves the production of complementary DNA (cDNA) of the RNA using a reverse transcriptase enzyme and random primers, which amplify all RNA within

Strain	Genotype	Description
W81 [162]	Δ*mmfL* Δ*mmfH* Δ*mmfP*	No MmF signalling molecules produced, hence no methylenomycin production
W89 [165]	Δ*mmyR*	Overproduces MmF signalling molecules and methylenomycin A
W95 [162]	Δ*mmyR* Δ*mmyD*	MmF signalling molecules produced, but no methylenomycin or intermediates
M145 [165]	SCP1- SCP2-	Methylenomycin cluster absent, no MmF, MmA or intermediates.

Table 5.3: Genotype and description of *S. coelicolor* strains used in RT-PCR assay. M145 is included for use as a positive control only.

the cell lysate [169]. This mixture of cDNA of different expressed genes can then be amplified with primers specific to the genes under scrutiny to determine whether they are indeed expressed in the cellular system. In order to ensure that no chromosomal DNA is present, which would lead to false positives, the cell lysate must first be processed to remove this DNA. Also, extra sterilisation procedures must be followed to remove any trace amounts of RNase which would disrupt results. In this instance, Life Technologies RNaseZap® was used to wash all equipment and work areas.

In order to investigate the expression of *mmyJ* and *mmr* genes, cultures of *S. coelicolor* strains W81, W89 and W95 (see Table 5.3) were grown and cDNA was obtained. It was expected that if the MmFs were binding ligands, then an increase in *mmr* and *mmyJ* expression would be seen in strains W89 and W95. However, increased expression of *mmr* and *mmyJ* in W89 only would indicate that just methylenomycin A and/or its intermediates bind to MmyJ, causing dissociation of the MmyJ:DNA complex.

Oligonucleotide pairs 16-18 from Section 8.1.3 were used to amplify the *hrdB*, *mmr* and

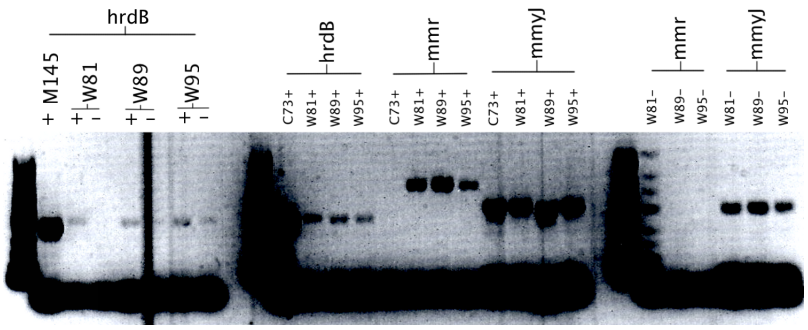


Figure 5.18: 1% agarose gel with results of RT-PCR assay. + and - indicate presence or absence of cDNA, with strains of origin indicated. Ladder was included but has overpowered the camera due to the high acquisition time required for PCR products to be visible. Dark bands at the bottom of each lane corresponds to primers. Leakage of ladder into W81- *mmr* lane is noted.

mmyJ genes present in the cDNA from each of the three W strains. *hrdB* was included as a positive control, as all strains would express this gene. Other positive controls were also included, using genomic DNA from strain M145 [165] with the *hrdB* primers and C73 genomic DNA with the *mmr* and *mmyJ* primers. Negative controls were also run using extracts that did not contain cDNA, to ensure no chromosomal DNA was present.

The result of this assay can be seen in Figure 5.18. Initial observations include the presence of bands corresponding to the *hrdB* negative control for W95, along with all three negative controls using the *mmyJ* primers. This indicates that there is some chromosomal DNA present in at least the W95 cDNA, if not all three strains, unfortunately invalidating any results. It is also apparent that the C73 positive control for the *mmr* primers failed, although the same primers did amplify the gene in all three experimental cDNA samples.

While this method shows promise, it was decided that other methods would be more efficient, especially in light of the previous EMSA results which indicate MmyJ does bind to DNA in such a way that transcription of *mmr* and *mmyJ* is repressed. Hence, this work was not taken any further.

5.3.2 Luciferase Reporter System

A reporter system has previously been developed for use in bacteria with a high GC content utilising the bioluminescent protein luciferase [170]. In this system, a promoter containing a transcriptional repressor binding site can be inserted upstream of the *luxCDABE* operon, such that the binding and release of the transcriptional repressor can be monitored by the level of bioluminescence, corresponding to the amount of luciferase produced. The pMU1 plasmid, containing this system as well as an apramycin resistance cassette, can be seen in Figure 5.19. This plasmid had previously been obtained from the Nodwell group at the University of Toronto for use investigating the binding of MmfR to the *mmfL* promoter region. As such the plasmid was easily accessible for this current work, with the intention of manipulating the system to mimic the expression of *mmr* under the control of MmyJ. Once constructed, the plasmids containing the *mmr* promoter would then be conjugated from *E. coli* into a *Streptomyces coelicolor* M145 host, which does not contain the SCP1 linear plasmid on which the methylenomycin gene cluster is located [165].

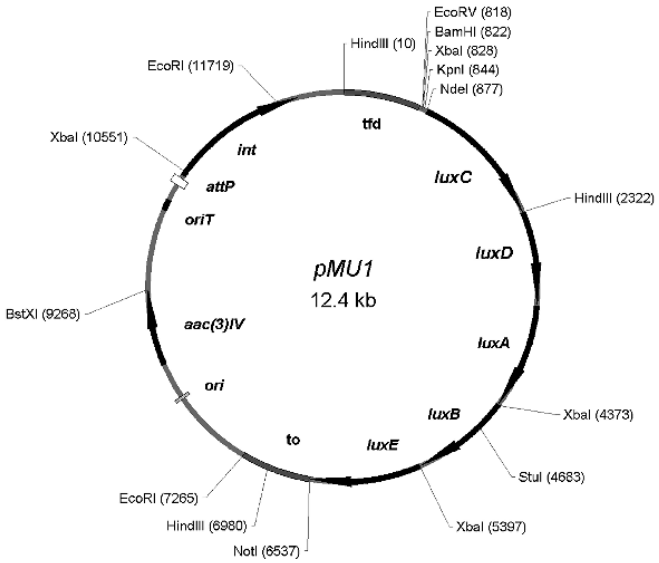


Figure 5.19: Plasmid pMU1 containing luciferase system controlled by specific promoter that can be inserted upstream of *luxCDABE* operon. *aac(3)IV* is an apramycin resistance gene. *int* and *attP* are the integrase gene and attachment site of ΦBT1 phage, allowing insertion of this plasmid into *Streptomyces* chromosomal DNA.

Primers were designed to amplify regions of C73 containing just the intergenic region between *mmyJ* and *mmr* (designated LmJ1) and also containing the intergenic region plus the *mmyJ* gene (LmJ2), as shown in Figure 5.20. These were designed such that the resulting PCR products would have the addition of a *Bam*HI restriction site at the *mmr* end and an *Eco*RV restriction site at the *mmyJ* end. In this way they would be the correct orientation to simulate *mmr* expression via the *lux* genes when inserted into the plasmid. Primer pairs 19 and

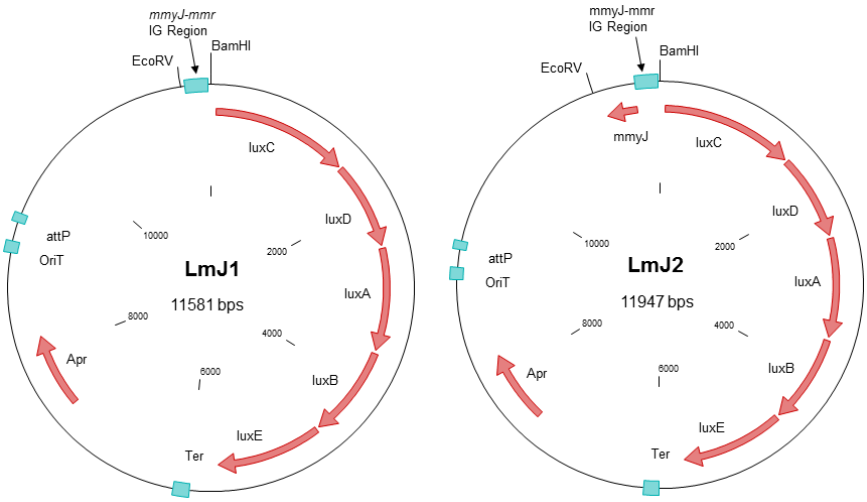


Figure 5.20: Plasmids LmJ1 and LmJ2, based on pMU1. Inserts were designed such that LmJ1 should express luciferase and LmJ2 should have the transcription of luciferase regulated by MmyJ binding to the promoter site. *mmyJ-mmr* intergenic region is orientated so as to simulate *mmr* expression, i.e. with the *mmr* promoter at the same end of the insert as the *Bam*HI digestion site.

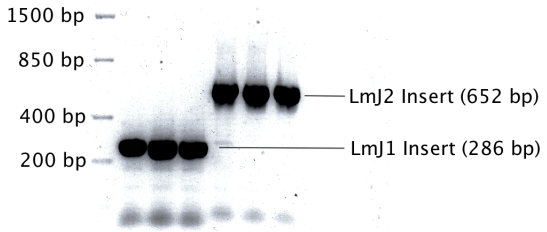


Figure 5.21: 1% agarose gel containing PCR products from amplification of LmJ1 and LmJ2 inserts. Each amplification was repeated in triplicate.

20 shown in Section 8.1.3 were used to create these inserts, which were amplified using Roche High Fidelity polymerase. The PCR products were then run on a 1% agarose gel as shown in Figure 5.21 to confirm they were the correct size. *Bam*HI and *Eco*RV were then added to the PCR products so as to digest the ends ready for ligation. The digested inserts were then gel purified.

Alongside this amplification and purification of the insert sequences, cells containing plasmid 11NY were grown in overnight cultures. This plasmid is a pMU1 derivative containing an 838 bp *Eco*RV/*Bam*HI insert designed to investigate MmfR binding. The 11NY plasmid was extracted and digested using *Bam*HI and *Eco*RV to remove this previous insert, giving a linearised version of pMU1 with digested *Bam*HI and *Eco*RV sites at each end. The result of this digestion is shown in Figure 5.22, demonstrating that the 11NY insert was successfully removed. The digested plasmid was then purified from the gel.

Ligation of the LmJ1 and LmJ2 inserts into the linearised pMU1 backbone was then attempted in triplicate with NEB T4 ligase. As *Eco*RV digestion leaves a blunt end, ligation was left to continue overnight in accordance with the manufacturer's recommendations. Colony number was small, with only 5 colonies appearing. Cells were taken from each colony to innoculate 10 mL liquid cultures and were grown overnight at 37 °C with shaking at 180 rpm. The plasmids were then extracted and sent for sequencing, however all came back negative, with the

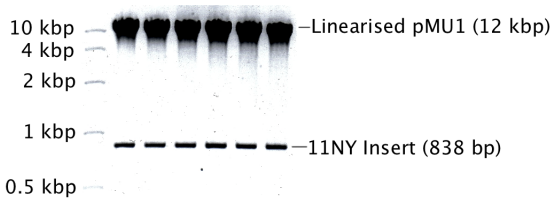


Figure 5.22: 1% agarose gel showing removal of 11NY insert from pMU1 plasmid after digestion with *Bam*HI and *Eco*RV.

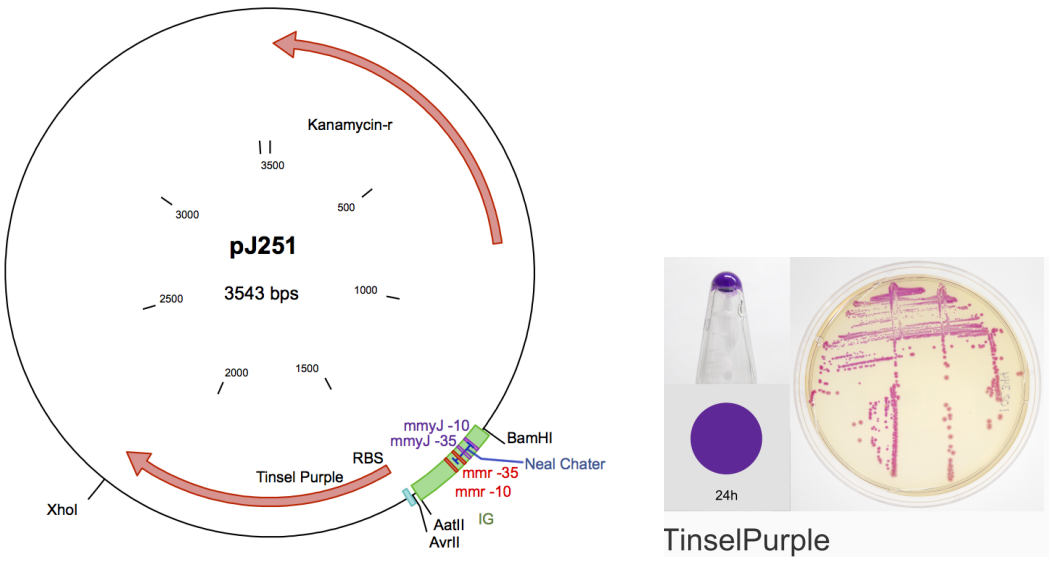
pMU1 plasmid apparently recircularising without the inserts in all cases. The transformation was repeated with freshly digested plasmid and an altered insert:plasmid molar ratio of 20:1 in an attempt to compensate for the small size of the insert. However, this time there was no colony growth at all.

The experiment was then repeated using Quick Ligase, again from NEB [171]. Also, this time only 0.025 mg/mL apramycin was added to the LB plates in order to reduce the amount of stress on the freshly transformed cells. In this instance there was good colony growth, and cells were picked from 48 colonies for overnight liquid cultures; 24 each of LmJ1 and LmJ2. These 10 mL cultures contained 0.05 mg/mL apramycin as before, as it was thought that the cells would be resistant to this level of antibiotic, so as to act as a further screen against unsuccessful transformations. Indeed, of the 48 overnight cultures, only 8 each of LmJ1 and LmJ2 grew, and so these then had their plasmids extracted for analysis. Each of these possible hits was then sent for sequencing. In this case primer pair 21 was used, which had been specifically designed to check for any inset between the *Bam*HI and *Eco*RV restriction sites, and had been successfully used previously by other group members investigating 11NY insertion. Unfortunately, sequencing again indicated no positive results, with no insert present in any plasmid.

At this point, this line of investigation was halted in favour of using a specially designed reporter system for use in *E. Coli* instead of *Streptomyces*, as described in the next section.

5.3.3 'Tinsel Purple' Reporter System

After the above negative results, it was decided that it would be more efficient to purchase a custom reporter system already containing the *mmr-mmyJ* intergenic region. As such, a Tinsel Purple reporter system, named pJ251, was purchased from DNA 2.0. This contained the intergenic region and a kanamycin resistance cassette as seen in Figure 5.23. In this case, it was intended that the plasmid would be transformed into BL21 star *E. coli* cells; both empty and already containing the pET151 His_6-MmyJ expression plasmid. As such, it was expected that on the addition of IPTG, expression of the T7 RNA polymerase would be derepressed and the cells containing just pJ251 would express Tinsel Purple, while those containing both pJ251 and pET151 His_6-MmyJ would not due to MmyJ repressing transcription from the *mmr* promoter.



(a) (b)

Figure 5.23: (a) Plasmid pJ251, designed to express a purple protein via the *mmr* promoter. The green region is the 218 bp *mmr-mmyJ* intergenic region, with -10 and -35 sites indicated for both promoters. The region labeled 'NealChater' is the 13-1-13 inverted repeat indicated in [100], since shown to be the MmyJ binding site. (b) Image taken from DNA 2.0 website showing colour of expressed protein after 24 hours in LB media [172].

The pJ251 plasmid was transformed into empty BL21 star cells, with resulting colonies being used to inoculate 10 mL cultures of LB media containing 0.1 mg/mL kanamycin and 1 mM IPTG. While the cells grew well, turning the culture distinctly cloudy, no Tinsel Purple was expressed. A sample of pJ251 was at this point sent for sequencing using primer pair 4, and was shown to contain the complete *mmr-mmyJ* intergenic region upstream of the gene encoding for Tinsel Purple as designed.

One possible explanation for this failure in expression stems from the initial design consultation, where it was decided that the kanamycin resistance cassette should be antisense to the expression gene for Tinsel Purple. It was mentioned that this may reduce expression, but it was desired so as to reduce the risk of read-through from the resistance cassette to the expression system, and hence prevent false positives. As there are unique restriction sites at either end of the expression system, it would be possible to design primers to amplify the region between *Xho*I and *Bam*HI sites such that it could be reinserted the other way around in order to test this hypothesis. Another possibility is that the *mmr* promoter may not be compatible with *E. coli* RNA polymerase, and so the insertion of a strong *E. coli* promoter site upstream of the inserted intergenic region may be required. In this way, the *mmr* promoter would be circumvented, but

MmyJ binding could still be investigated as it would obstruct the progression of the RNA polymerase along the DNA sequence. However, there has not been time to realise either of these experiments, and so further work must be done in order to attempt this and carry on the *in vivo* aspect of the investigation into the functionality of MmyJ.

6 Structure

While a structure based on homology modelling has already been generated and presented in Section 2.3, a 3D structure determined by experimental methods would give a more reliable insight into the properties of MmyJ.

6.1 Secondary Structure Analysis by Circular Dichroism

As already proven in Section 4.1.2, MmyJ has been found to have a high degree of α-helical characteristics, as befitting an ArsR family protein. Subsequent work has stemmed from this assertion, however the exact degree of α-helical character has not previously been alluded to.

A Circular Dichroism (CD) spectrum of His_6-MmyJ in sodium phosphate buffer, obtained as described in Section 4, was uploaded to the DichroWeb [173, 174, 175] server for analysis with SELCON3 [176, 177], CONTIN/LL [178, 179] and CDSSTR [180, 181, 182] algorithms, as recommended in [182]. In all cases, Reference Set 4 [182] was used, which is optimised for data between 190 and 240 nm, corresponding to the range containing characteristic peaks for α-helical and β-sheet secondary structure, as discussed in Section 4.1.1.

6.1.1 SELCON3 Analysis

SELCON3 is the latest iteration of the SELCON algorithm, which is a self consistent method for the determination of secondary structure of proteins from CD data [176]. It works by including the experimental data in the basis set, which it then uses to create an initial guess of the protein secondary structure. This initial guess is based heavily on the reference spectrum found to be most similar to the experimental data, and is used to predict a second approximation of the predicted structure. This then replaces the initial guess in the next iteration of the algorithm, which continues until it converges on a model most suited to the experimental data. Figure 6.1 shows the resulting modelled spectrum alongside the experimental spectrum submitted. From the plot of deviation, it can be seen that this model does not perfectly replicate the experimental data. However, as the three algorithms featured are optimised for different aspects of secondary structure, it is recommended to perform all three analyses and evaluate the results together [182].

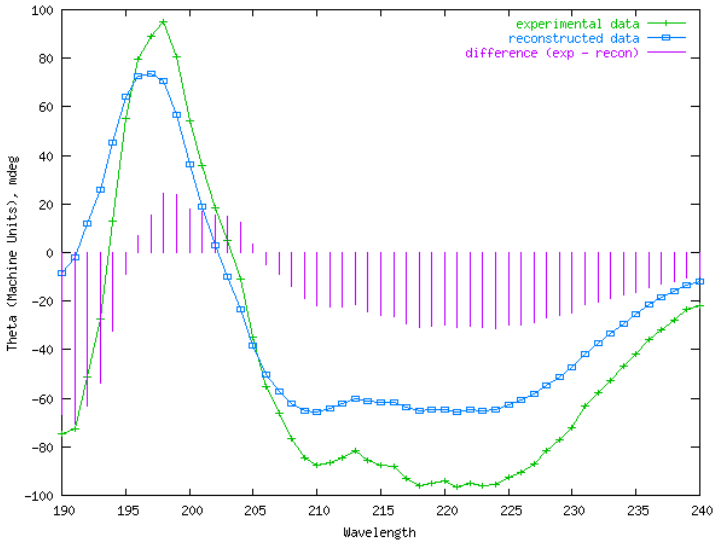


Figure 6.1: Comparison of CD spectrum expected from SELCON3 fit of predicted secondary structure of His_6-MmyJ to normalised experimental data. Deviation is shown as purple lines. Model predicts 57% helical, 5% strand, 19% turn and 20% disordered secondary structure.

6.1.2 CONTIN/LL Analysis

CONTIN/LL is a modification of the CONTIN algorithm, which uses a ridge progression procedure to minimise the sum of differences between the experimental spectrum and the reference set, with weighting added according to a linear combination of iterations of the algorithm [178, 179]. The CONTIN/LL modification reduces the reference set to exclude proteins with a high RMSD when compared to the experimental data, thus reducing the

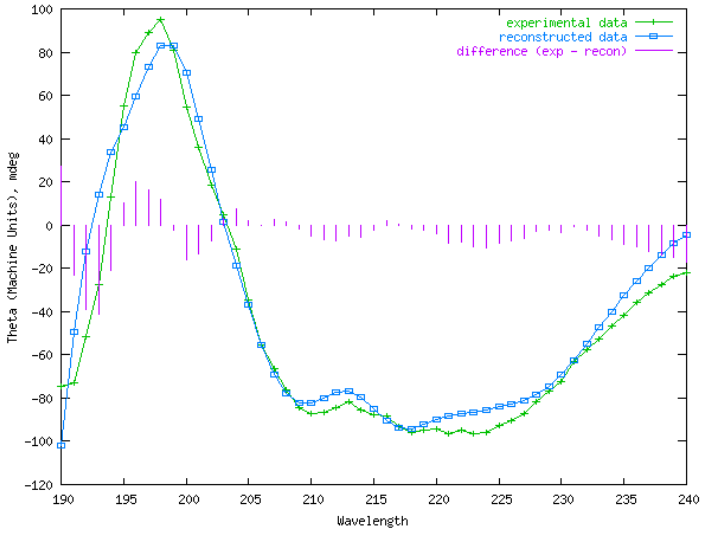


Figure 6.2: Comparison of CD spectrum expected from CONTIN/LL fit of predicted secondary structure of His_6-MmyJ to normalised experimental data. Deviation is shown as purple lines. Model predicts 68% helical, 6% strand, 13% turn and 13% disordered secondary structure.

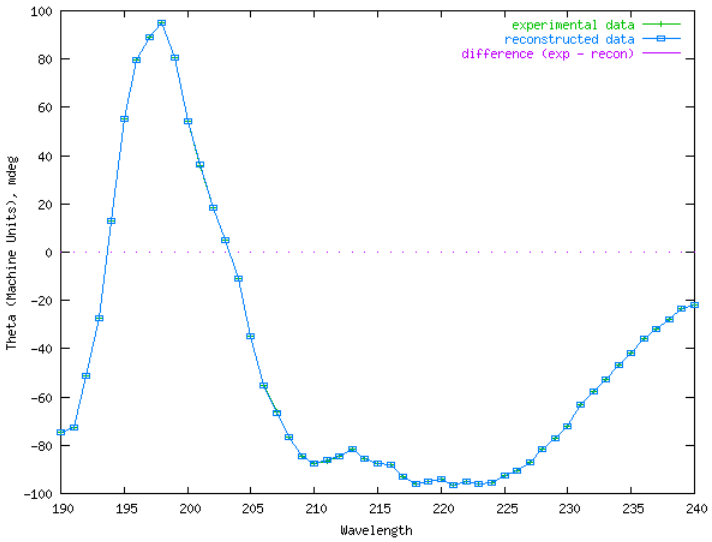


Figure 6.3: Comparison of CD spectrum expected from CDSSTR fit of predicted secondary structure of His_6-MmyJ to normalised experimental data. Deviation is shown as purple lines, which in this instance are nearly dot-like. Model predicts 53% helical, 21% strand, 9% turn and 18% disordered secondary structure.

expected error in the final model [182]. Figure 6.2 shows the modelled spectrum for this algorithm compared to the experimental result, and it can instantly be seen that it offers a better fit that the SELCON3 algorithm.

6.1.3 CDSSTR Analysis

The final method used to fit the experimental data is CDSSTR, which randomly selects just 8 proteins from the reference set, with iterations of different combinations of these reference proteins compared to the experimental data so as to satisfy the algorithm's selection rules [182]. Figure 6.3 displays the resulting approximated spectrum. It can be seen in this instance that the fit is near perfect, with no observable deviation between it and the experimental data. It should be noted that due to the random nature of the reference selection, this result may not be stable to attempts to reproduce it [182]; however the analysis presented here was repeated and there seemed to be no noticeable deviation.

6.1.4 Comparison of Models

While an initial visual comparison implies that CDSSTR is the best algorithm and SELCON3 is the worst at modelling His_6-MmyJ (assuming the homology model is accurate), the resulting modelled characteristics need to be examined more closely and compared to other models

Model	% Helical	% Strand	% Turn	% Disordered
SELCON3	57	5	19	20
CONTIN/LL	68	6	13	13
CDSSTR	53	21	9	18
Mean of CD Models	59	11	14	17
Phyre2	62	10	-	24 (28)

Table 6.1: Comparison of secondary structures predicted by previously discussed models based on CD data as well as the new Phyre2 modelling of His_6-MmyJ. It should be noted that 4% of the residues in the Phyre2 model are unassigned, and have therefore been added to the % Disordered column to give the value in brackets.

of predicted features; namely, in this instance, the homology model of MmyJ discussed in Section 2.3. It should be noted that this model is of MmyJ as it occurs *in vivo*, rather than the His tagged fusion protein on which the above analyses were carried out. As such, a new Phyre2 [112] model was created[8] with the addition of the His tag and V5 epitope [124].

The comparison between these four models is shown in Table 6.1. It can be seen that while no individual model matches the homology model, the mean across all three models is in good agreement with the homology model when considering the helix and sheet content of His_6-MmyJ. There is less obvious agreement between the % Turn and % Disordered, but that is largely because Phyre2 does not assign residues as being part of a turn structure, and hence many of the residues assigned as disordered may in fact be parts of an unrecognised turn. For example, one of the residues between helices α3 and αR in the new Phyre2 model is classed as disordered, despite it being one of the three residues that comprise the turn in the HTH motif. As such, if the mean % Turn and % Disordered values from the CD models are summed together, the total value of 31 can be seen to be close to the adjusted Phyre2 value of 28. In this way the Phyre2 model is somewhat validated by the CD analyses; at least as far as relative percentage composition of helical, strand and turn/disordered is concerned.

6.2 Nuclear Magnetic Resonance Spectroscopy Characterisation

In an attempt to obtain more complex structural information, Nuclear Magnetic Resonance (NMR) spectroscopy was performed with ^{15}N labelled protein samples.

[8]The new Phyre2 model is not shown here as it deviates from the previous model only in the addition of an extra α-helix within the V5 epitope, which is only relevant when comparing to the CD data.

6.2.1 Quantum Mechanical Theory - A Brief Overview

The basis of NMR lies in the possession of intrinsic angular spin momentum by all nuclei. While orbital angular momentum has both a classical and quantum mechanical interpretation of an electron orbiting a nucleus, spin angular momentum can not be thought of in terms of classical physics, and can only be understood through quantum theory.[9] Subatomic particles can be classified as either bosons (such as photons) or fermions (such as neutrons, protons and electrons) by their intrinsic integer or half-integer spin quantum number respectively. This is denoted m, and is given in units of $\hbar = h/(2\pi)$, where h is Planck's Constant [183]. As such, atomic nuclei can have a range of spin values, calculated as the sum of spin values of their constituent protons and neutrons. These spin values give the magnitude of the spin angular momentum, usually represented by $I = |m|$. However, due to the Heisenberg Uncertainty Principle, only the total magnitude and one component, typically taken to be that in the z direction, can be known definitively. This can then be scaled by the gyromagnetic ratio γ, a unique nuclear property that varies with isotope, to give the magnetic moment μ as follows [184]:

$$\mu_z = \gamma I_z \tag{6.1}$$

This is the source of the NMR signal, and so it can be seen that in the case of nuclei with a zero spin there is no magnetic moment. This includes ^{12}C, the most abundant isotope of carbon, hence NMR can not be performed on these nuclei. Another isotope of carbon, ^{13}C, has a spin value of $\frac{1}{2}$, and, as such, can be detected by NMR [185]. However, this isotope only has a natural abundance of 1%. This is enough to generate a useable signal for small molecule NMR, but for complex molecules such as proteins, labelling strategies must be employed to increase the presence of ^{13}C.

Some nuclei have a spin number whose magnitude is greater than $\frac{1}{2}$, and so their electromagnetic profile contains higher order functions, such as a quadrupole moment [184]. This adds extra complexity to the acquisition and processing of NMR data, and so labelling strategies are again employed to enhance the presence of half-spin nuclei. For example the most abundant isotope of nitrogen is ^{14}N, for which $I = 1$, leading to quadrupolar interactions. Molecules

[9] The name spin angular momentum is in fact misleading, and originates from early theories when it was thought to stem from the spin of an atom on its axis. This has since been proven not to be the case.

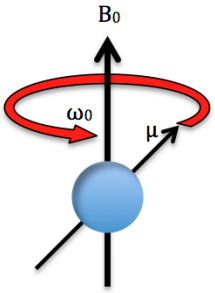


Figure 6.4: Illustration of precession of a magnetic moment μ around an external magnetic field B_0 at frequency ω_0.

are therefore often enriched with ^{15}N, for which $I = \frac{1}{2}$ [185], allowing NMR experiments to be performed without extra complication.

In order to understand an NMR experiment, it is also necessary to be aware of the interaction of μ with an external magnetic field, B_0. All nuclei possessing a non-zero magnetic moment will align with an external magnetic field. However, as it is impossible to know all components of I, and hence μ, the dipole moment cannot completely align with B_0, and as such it precesses around the magnetic field as demonstrated in Figure 6.4. The frequency of precession is defined by the gyromagnetic ratio and magnetic field strength as follows, and so is different for all isotopes [184]:

$$|\omega_0| = \gamma |B_0| \tag{6.2}$$

This precession frequency, ω_0, is called the Larmor frequency and is a characteristic property of each isotope in a given magnetic field. As such, NMR spectrometers are often named by the Larmour frequency of a ^{1}H nucleus in the magnetic field produced, for example a 14.1 T magnet would cause ^{1}H nuclei to precess at approximately 600 MHz, and so this is defined as the field strength of such a spectrometer.

In an NMR spectrometer there is also an additional, weaker magnetic field B_1, which is applied perpendicular to B_0. This field oscillates at a tuneable frequency ω_{rf}, corresponding roughly to the frequency of radio waves, such that the field at any given time can be defined as [184]

$$\mathbf{B_1} = |B_1| \cos(\omega_{rf} t + \phi) \tag{6.3}$$

where ϕ is the phase of the oscillation. When $\omega_{rf} = \omega_0$ for a given isotope (defined as the

resonance condition), the magnetic moment nutates about the B_1 pulse, out of alignment with B_0. As the magnetic moment returns to equilibrium (i.e. aligned with B_0), its precession is recorded as a Free Induction Decay (FID), which can be Fourier Transformed to give the frequency corresponding to the nuclei perturbed by the B_1 pulse [186]. As this frequency is inherently dependant on the strength of the magnetic field used, it is typically quoted as a Chemical Shift in units of parts per million (ppm), usually denoted as δ. This allows the direct comparison of values obtained from different spectrometers, using the conversion as follows [184]:

$$\delta(\text{ppm}) = \frac{\omega_0 - \omega_0^{ref}}{\omega_0^{ref}} \tag{6.4}$$

where ω_0^{ref} is the Larmor frequency of the isotope being investigated, taken from a known reference sample, e.g. 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) or tetramethylsilane (TMS). It should be noted that NMR signals are inherently weak; the detected signal is actually the difference between nuclei which align with B_0 and those that align against it; a difference of only one nucleus in 10,000, as defined by Boltzmann statistics at standard temperature and pressure [184]. As such, multiple scans are typically carried out and summed together. As signal occurs at the same frequency each time whereas noise occurs randomly; summing N repeated experiments causes the signal to noise ratio to increase by a factor of $1/\sqrt{N}$ [184].

The power of NMR as a tool for structural characterisation of molecules comes from the fact that individual molecules do not necessarily experience the full B_0 field due to local shielding by electrons and other effects of intermolecular interactions and bonds [184]. By this fact, the resonance frequency of nuclei varies with the local environment, and as such a spectrum of frequencies can be obtained where each peak corresponds to a unique electronic environment. For simple molecules, this can be done for just one or two atomic species in order to determine the local environment of each atom and therefore solve its structure. If investigating multiple nuclei, a combination of pulse sequences can be used that allow interactions between atoms of the same (homonuclear), or different (heteronuclear), species to be explicitly measured [184]. This sort of data is usually displayed as a two dimensional topographic map indicating coupled nuclei. However, for larger molecules such as proteins, more complex experiments are also required, leading to three-, four- or five-dimensional data. This can be visualised as an ensemble

of topographic cubes and maps, allowing interactions to be tracked between molecules either directly bonded or close in space, from which interatomic spacing can be computed. Some experiments implemented in this work are two dimensional, however their design requires a deeper understanding of quantum mechanics that is beyond the scope of this thesis.

6.2.2 Protein NMR in the Solution State

Protein NMR in the solution state has led to the elucidation of many protein structures, however there are inherent limitations in this method. The main issue lies with the restraints of averaging the anisotropic interactions stemming from the orientation of interatomic bonds with respect to the external magnetic field, many of which have the general form [184]

$$C = C_{Aniso}\frac{1}{2}(3\cos^2\theta - 1) \tag{6.5}$$

where C_{Aniso} denotes the interaction due to anisotropic coupling of nuclei and θ denotes the angle between the inter-atomic bond and the external magnetic field. These interactions, when taken over an ensemble of molecules in a sample, lead to broadening of the peaks within the NMR spectrum, reducing resolution and sensitivity of the method. However, in the solution state these interactions are averaged by molecular tumbling (the rotation of molecules within the solution), such that they are effectively cancelled out as follows:

$$\int_0^{\pi} \frac{1}{2}(3\cos^2\theta - 1)\sin\theta\,\mathrm{d}\theta = 0 \tag{6.6}$$

Hence the anisotropic interactions tend to zero so long as molecular tumbling occurs within a shorter timeframe than the measurement of the FID. However, tumbling speed reduces as molecular size increases, and as such there is an upper limit in the size of molecule that can effectively be studied by solution NMR [184]. There are exceptions, as complex labelling strategies or pulse sequences can be implemented to raise this upper limit, but these add more expense, time and complexity to the experiments.

A 2D ^{1}H-^{15}N Heteronuclear Single Quantum Coherence (HSQC) experiment [187] was performed on a sample of MmyJ uniformly labelled with ^{15}N, prepared to a concentration of

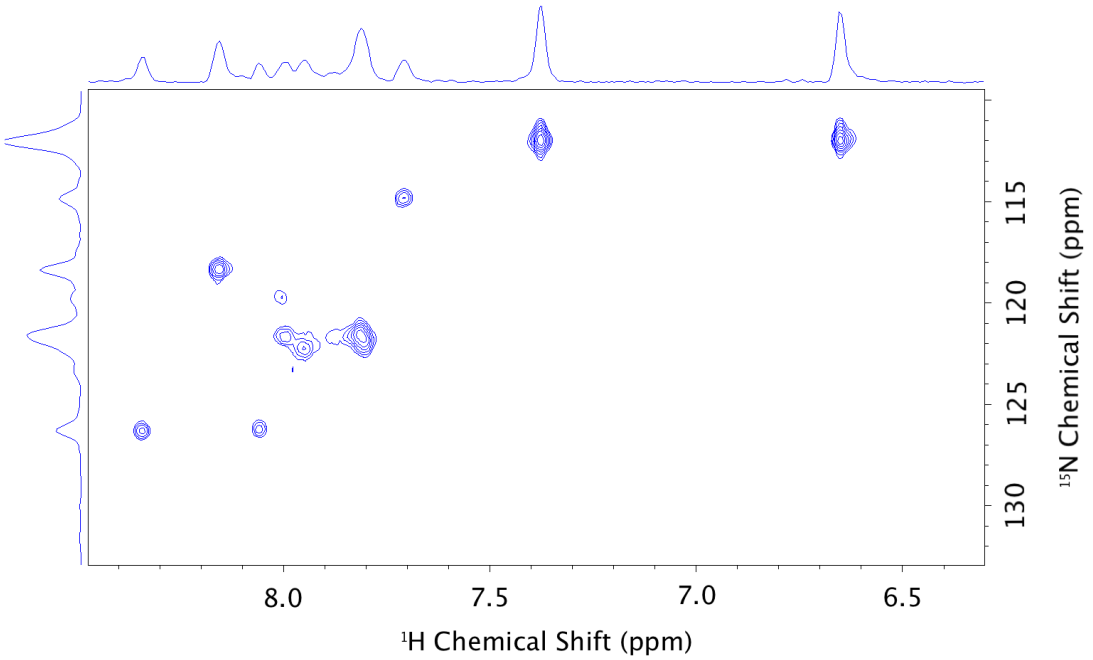


Figure 6.5: ^{1}H-^{15}N HSQC spectrum of uniformly ^{15}N labelled MmyJ in 50 mM Tris-HCl pH 7. Summed spectrum of 256 scans.

approximately 1 mM in 50 mM Tris-HCl pH 7 containing 10% D_2O and 0.1 mM DSS reference. The experiment comprised 256 scans, which were summed together to increase the signal to noise ratio as described above. The resulting spectrum can be seen in Figure 6.5. Ideally this spectrum would contain a peak corresponding to every N-H interaction within MmyJ, of which there should be at least 111 (one for each amine bond). It is apparent that this is not the case, as at most only 10 individual peaks can be identified. It was suspected that the pH was too high, and so pH was lowered to approximately 5.6 to reduce the rate of proton exchange between the backbone amides and water/deuterium in the solvent [188]. The spectrum for MmyJ in this altered solvent can be seen in Figure 6.6 where, in an effort to further increase signal, 896 scans were acquired and summed. However, it can be seen that while a couple more peaks are apparent, there is not a great deal of improvement upon the previous spectrum.

It was noted during sample preparation that aggregates were forming and dropping out of solution which, after attempts to observe the phenomenon via gel filtration, was found to be the direct result of removing MmyJ from an environment containing glycerol. Unfortunately, glycerol could not be retained in the solution used for NMR analysis as the increased viscosity would have much the same effect as if studying a much larger protein; the rate of molecular tumbling would be drastically reduced to such a point that an NMR spectrum would not

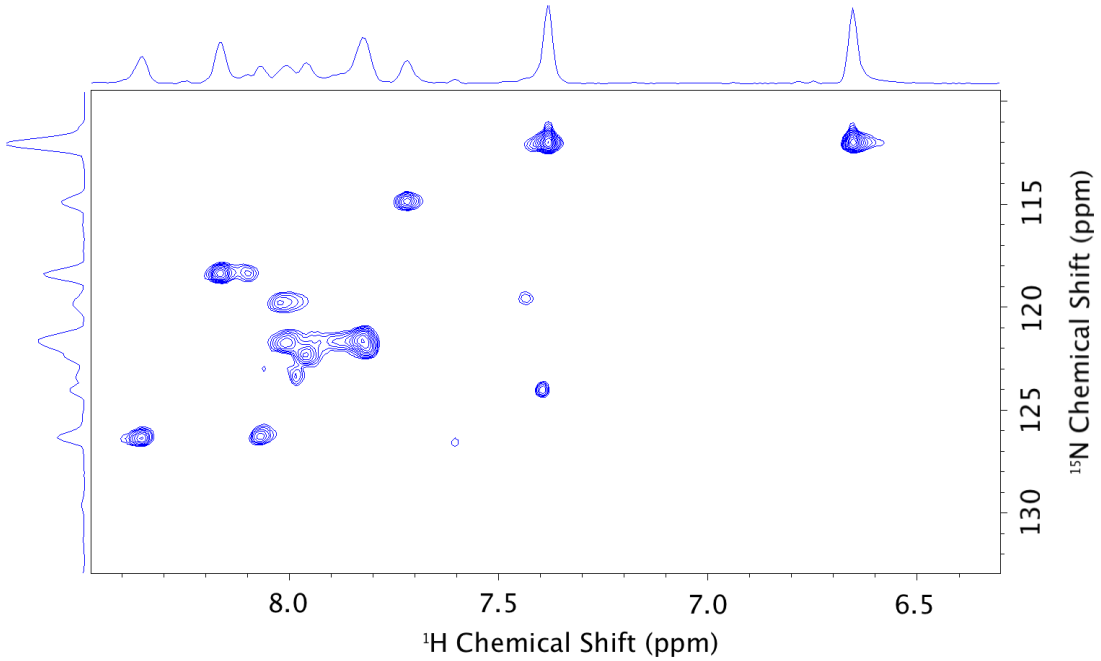


Figure 6.6: ^{1}H-^{15}N HSQC spectrum of uniformly ^{15}N labelled MmyJ in 50 mM Tris-HCl pH 5.6. Summed spectrum of 896 scans.

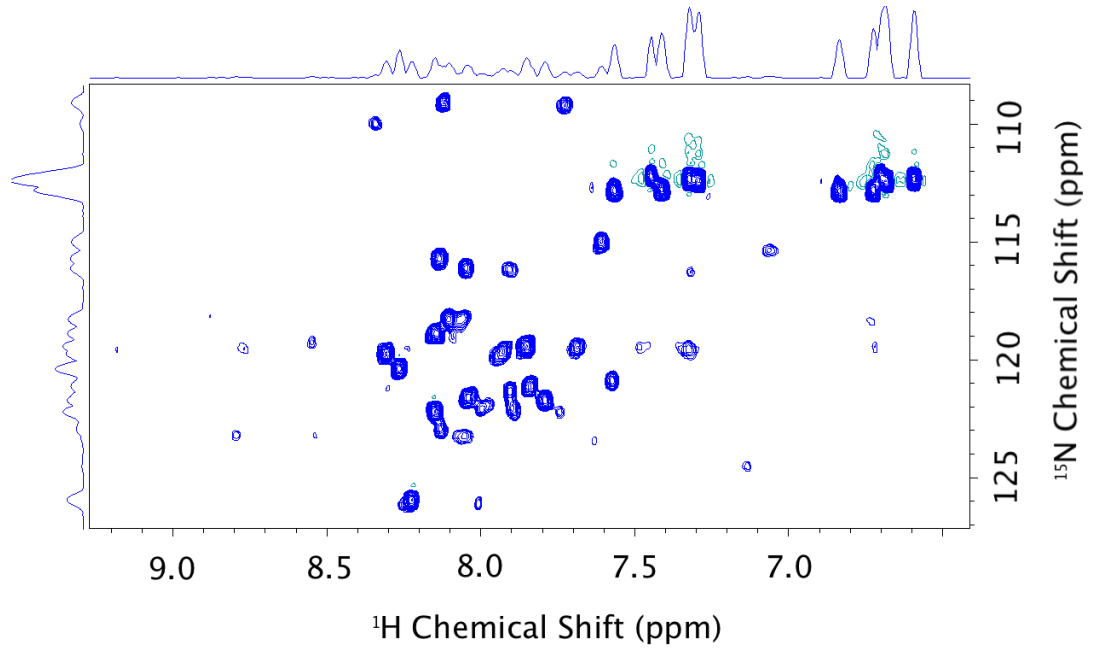


Figure 6.7: ^{1}H-^{15}N HSQC spectrum of uniformly ^{15}N labelled MmyJ in 50 mM H_2KPO_4 pH 7 with 50 mM arginine and glutamine added to improve stability. Summed spectrum of 616 scans.

be obtainable. An alternative was found in the addition of arginine and glutamine at a concentration of 50 mM in place of glycerol, as this has been shown to give an approximately 9 times increase in protein stability [189]. As well as this, phosphate buffer was used rather than Tris-HCl, containing just H_2KPO_4 at a concentration of 50 mM so that the arginine acted as a base in this solution.

Figure 6.7 displays the results of performing the same HSQC experiment as before using the new buffer at pH 7. It is immediately evident that there are many more peaks present, indicating that stability was indeed likely to be a contributing factor to the previous issue. However, there still are too few peaks present to attempt to assign them to residues within MmyJ. It was thought that the absence of further peaks was likely due to the size of the MmyJ dimer, as the upper limit on proteins observable via simple ^{15}N labelling is considered to be approximately 150 residues [190], whereas the MmyJ dimer would be the equivalent of a 222 residue protein. Full peak assignment would require at least a three-dimensional data set, and the data shown in Figure 6.7 required over 15 hours of acquisition time just for two dimensional data. As it was not thought that lowering the pH would improve signal drastically, it was decided that other experiments should be investigated. Uniform ^{13}C, ^{15}N labelling was attempted to reduce acquisition time, but unfortunately protein overproduction failed when attempted. Also, other solution state methods such as Transverse Relaxation-Optimised Spectroscopy (TROSY) experiments were considered, but, as this would require deuteration, it was decided that attempting solid state NMR would be a better alternative.

6.2.3 Solid State NMR

As previously mentioned, the main limitation on the size of molecule that can effectively be characterised by solution state NMR is due to the slowing down of molecular tumbling to timescales longer than the acquisition of the FID. In the solid state, there is no molecular tumbling and so there is no averaging of the anisotropic interactions at all, leading to catastrophic line broadening with stationary samples [184]. However, by spinning the sample at high speed, molecular tumbling can be approximated such that anisotropic interactions are again averaged out, with increasing speed used to replicate increased molecular tumbling. This is not a perfect solution, however, as the interactions that lie along the axis of rotation are not

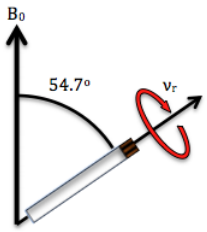


Figure 6.8: Illustration of so-called Magic Angle Spinning (MAS), with rotor containing sample rotating at 54.7° to B_0. The spinning frequency ν_r must be sufficiently large so as to replicate molecular tumbling.

averaged, leaving some anisotropic contribution present in the recorded signal. The solution to this lies in Equation 6.5, where it can be seen that if the $3\cos^2\theta - 1$ term in brackets is equal to zero, the contribution to anisotropic coupling along that specific angle will also be zero. This is easily solved as follows

$$\theta_M = \cos^{-1}\frac{1}{\sqrt{3}} = 54.7^{\circ} \tag{6.7}$$

where θ_M is classed as the so-called magic angle. In this way, if the sample is spun[10] at sufficient speed at an angle of 54.7° to the external magnetic field, as illustrated in Figure 6.8, then all anisotropic effects are averaged out with the exception of those along the axis of rotation, which also go to zero as an artefact of quantum mechanics [191].

In order to prepare a sample for solid state NMR, a newly developed method was utilised, wherein an almost saturated solution of protein is spun in an ultracentrifuge, resulting in forces approaching $10^6 \times g$, causing a protein concentration gradient across the sample. If concentrated enough, this gradient will lead to super-saturation and precipitation, which can be collected directly into the NMR rotor. It has been shown that this preparation method results in enough short range order for NMR data to be acquired, comparable to as if the sample was crystalline [192].

A solution of MmyJ in 50 mM phosphate buffer containing 10% glycerol[11] was concentrated to approximately 1 mM. This was then added to a vessel with a funnel leading to a 1.3 mm solid state NMR rotor and spun in a Beckman SW41 Ti Rotor at 25,000 rpm, corresponding to a force of approximately $100,000 \times g$. A precipitate was observed to form within the rotor, and a simple one dimensional pulse sequence was used in an effort to observe cross polarisation (CP),

[10] Due to the required absence of any mechanical parts, spinning is achieved via high pressure air flow over curved flanges built in to the sample rotor.

[11] As molecular tumbling is not an issue in solid state NMR there is likewise no issue regarding the retention of glycerol in the sample buffer.

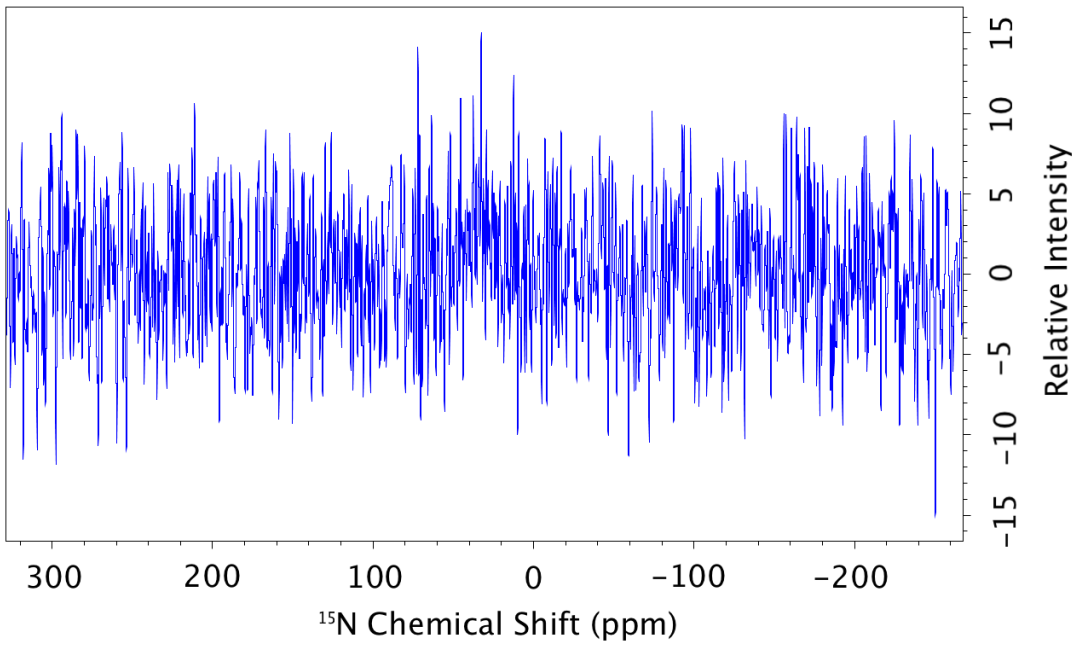


Figure 6.9: ^{1}H-^{15}N CP spectrum of uniformly ^{15}N labelled MmyJ precipitated from 50 mM phosphate buffer containing 10% glycerol. Summed spectrum of 2048 scans.

i.e. the transfer of magnetisation [193], between ^{1}H and ^{15}N nuclei, with magic angle spinning (MAS) carried out at 60 kHz. The resulting spectrum, summed over 2048 scans, can be seen in Figure 6.9. It is apparent that even with this many repeated spectra summed together, there is no visible spectrum demonstrating ^{1}H-^{15}N interactions. The sample preparation was repeated with fresh MmyJ, and again the same results were observed. It is thought that even though the MmyJ dimer was massive enough to form a precipitate, it may have been too small for said precipitate to form cohesively, with short range order present to the degree required to generate a detectable NMR signal.

Further solid state NMR experiments were planned using crystallised ^{15}N MmyJ (see Section 6.3.2 below for growth conditions), however the yield of crystals was insufficient to be of use in NMR. As such, X-ray diffraction of crystallised MmyJ became the main focus for determining its structural characteristics, as discussed in the remainder of this chapter.

6.3 X-Ray Diffraction

Further to the above attempts at NMR, X-Ray Driffraction was also utilised in an attempt to solve the structure of MmyJ.

6.3.1 A Brief Introduction to Crystallography

Diffraction occurs when a wavefront bends as it passes through a gap, with the most noticeable effects occurring when the wavelength is of a similar magnitude to the dimension of the gap. If more than one gap is present, then a distinct diffraction pattern is observed, formed of constructive and destructive interference between the multiple new wavefronts [194]. The typical classical example of this is the Young's Double Slit experiment, where light is shone through two adjacent slits on the order of 100 nm wide and onto a screen. On the screen a repeating pattern of light and dark bands is visible where the wavefronts interact [195]. The properties of this pattern, such as the width of the bands and distance between them, can be measured and, along with known information such as the distance between the slits and the screen, can be utilised to calculate the width and spacing of the slits. This principle also works for more complex two dimensional patterns, although the amount of computation involved obviously scales with the complexity of the pattern.

The theory holds for all wavelengths of light, and so X-rays can be diffracted by passing through gaps on the order of 0.1-100 Å. As the typical interatomic distances in molecules and ionic compounds fall within this range, X-rays can be tuned to approximately match these distances, allowing diffraction patterns to be observed. This type of diffraction, where a pattern arises through constructive and destructive interference of waves reflected from planes in a crystal lattice of atoms or molecules, is called Bragg Diffraction and is based upon Bragg's Law:

$$2d \sin\theta = n\lambda \tag{6.8}$$

where d is the distance between crystal planes, θ is the angle between the crystal plane and incident light beam that maximises constructive interference, n is a positive integer and λ is the wavelength of the incident wave [196].

When diffracting from a three dimensional crystal lattice, the above equation is insufficient on its own, and so theory has been developed to incorporate a reciprocal lattice, containing coordinates corresponding to the diffraction pattern [197]. This, along with Bragg's Law, was incorporated into a geometrical construction called an Ewald Sphere, from which the angle required to form a specific diffraction pattern from a crystal lattice can be calculated [198, 199,

197]. In this manner, the unit cell of a crystal may be calculated and defined. This is the smallest repeating volume which is then repeated through symmetry throughout the crystal; hence if a crystal is pure, then the contents of the unit cell are mirrored throughout its volume, and can be said to contain the structural information of the entire crystal [200]. In this way, it can be seen that if the unit cell is defined as containing a protein, and the protein is suspended within a crystal lattice such that the unit cell is repeated over and over, then the resulting diffraction pattern can be solved to give the structure of the unit cell, and so the protein itself [197].

Protein crystals are typically formed by either sitting or hanging drop methods. In both instances a small drop containing protein in solution and crystallisation buffer, also referred to as mother liquor [201], is sealed in an air tight well or compartment along with a large reservoir of the mother liquor, either directly below (hanging drop) or adjacent to (sitting drop) the protein solution. Vapour diffusion from the droplet to the reservoir causes the concentration of both protein and any added precipitants to increase slowly, such that eventually crystals start to form [200]. These can then be left to grow; however, if the well is not completely air tight then the reservoir and droplet will both dry out, usually destroying the crystals [200].

6.3.2 Crystallisation Screen

As proteins are a broad and diverse class of molecules, there is no one buffer solution that will lead to the correct conditions for crystallisation for all proteins. As such, screening assays are used in order to determine the conditions required for a specific protein to crystallise. In this work, four such 96 component screens were investigated:

- Morpheus - optimised to include ligands and buffers that were noted to have a high occurrence in crystallisation conditions listed alongside solved structures on the Protein Data Bank (PDB) [202].
- ProPlex - designed by screening over 600 conditions from published crystallised protein-protein complexes and optimising the mixtures of significant additives [203].
- JCSG+ - an updated extension of the JCSG screen, which itself contains the best 67 of 480 previously published screen conditions [204].

- PACT - a systematic test of pH, anions and cations with polyethylene glycol (PEG) added as a precipitant [204].

In all cases, cleaved MmyJ was purified and concentrated to approximately 10 mg/mL in 20 mM Tris-HCl pH 8.8 with 10% glycerol added by weight. Screening plates were then prepared as sitting drops containing 100 nL of MmyJ added to 100 nL of each screening condition, along with a 75 µL buffer reservoir for each sample. The screening plates were then sealed and left in a cool, dry room.

After approximately 3 weeks, it was noted that some of the screening conditions had led to the formation of small, rock-like crystals, while others had caused MmyJ to aggregate. It was noticed that PACT D3 (0.1 M Malic acid, MES and Tris + 25% PEG 1500, pH 6.0) contained large angular crystals, while PACT D5 (as D3, but pH 8.0) had small crystals. As this was the first pattern discovered, hanging drops were set up to screen around this condition, with a pH between 5 and 8 and PEG 1500 concentrations of 17.5-30% in 2.5% increments. These drops contained 1 µL MmyJ plus 1 µL screen condition suspended above a 200 µL reservoir. Crystals were observed to grow after several weeks, with an example shown in Figure 6.10, but unfortunately the in-house X-ray diffractometer was out of service for a prolonged period, and so these hanging drops dried up before they could be investigated.

During the time when the diffractometer was out of use, the initial screening plates were checked regularly in order to identify further conditions warranting investigation once possible. After three months, photos were taken of all hits from the initial screen, which can be seen in Figure 6.11 (NB: a photograph of PACT D5 is not included as the crystals grew at the edge

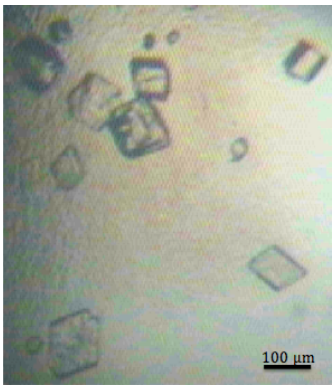
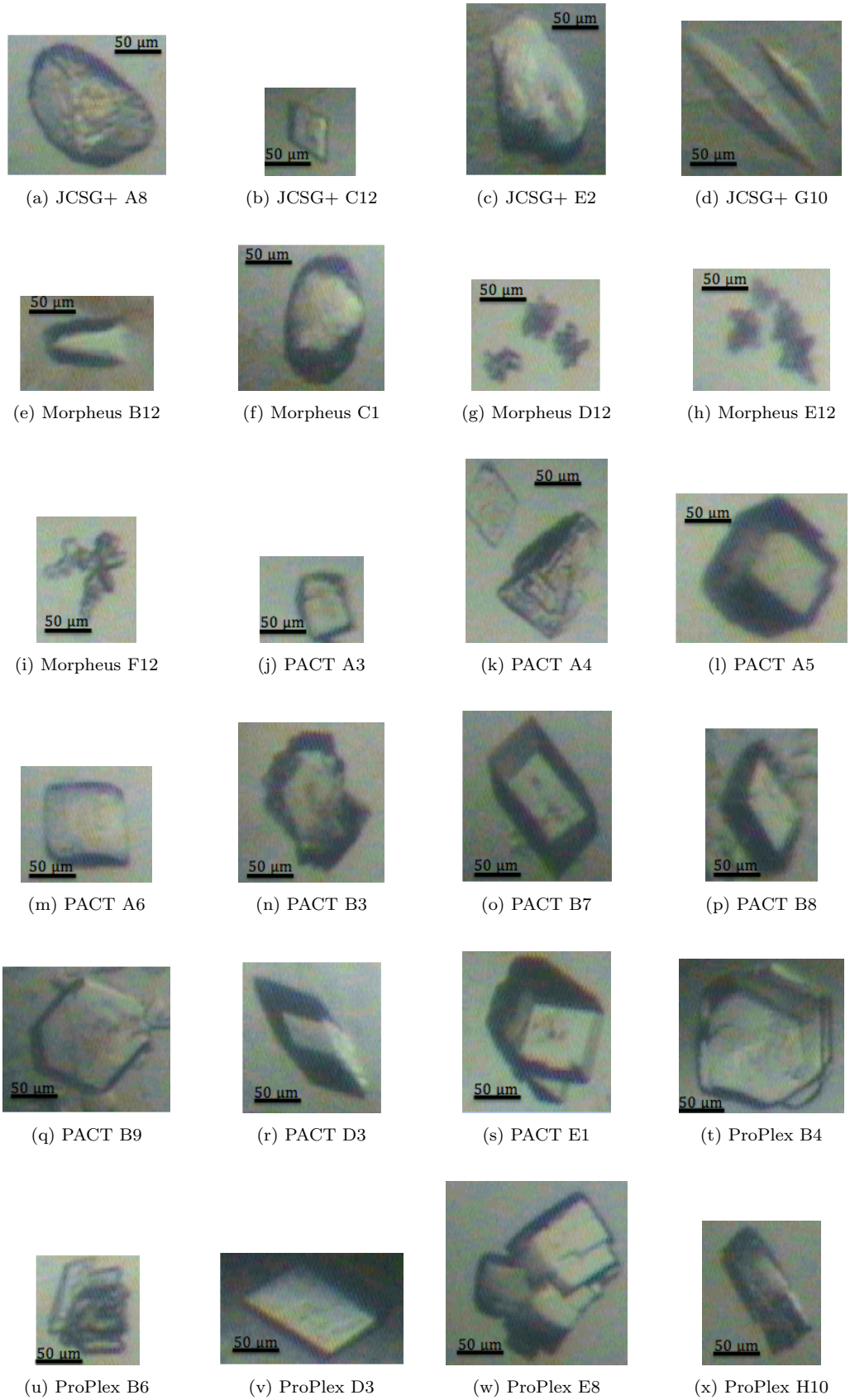


Figure 6.10: Photograph of crystals grown in 0.1 M Malic acid, MES and Tris + 17.5% PEG1500 pH 8.0. Taken approximately 2 months after hanging drop was set up.



(a) JCSG+ A8 (b) JCSG+ C12 (c) JCSG+ E2 (d) JCSG+ G10

(e) Morpheus B12 (f) Morpheus C1 (g) Morpheus D12 (h) Morpheus E12

(i) Morpheus F12 (j) PACT A3 (k) PACT A4 (l) PACT A5

(m) PACT A6 (n) PACT B3 (o) PACT B7 (p) PACT B8

(q) PACT B9 (r) PACT D3 (s) PACT E1 (t) ProPlex B4

(u) ProPlex B6 (v) ProPlex D3 (w) ProPlex E8 (x) ProPlex H10

Figure 6.11: Photographs of crystals grown from original screens. Corresponding JCSG+, Morpheus, PACT and ProPlex conditions can be found in [202, 203, 204].

of the drop of solution and as such cannot be seen due to contrast issues). Upon comparing these hits to the conditions in which they grew, several other patterns were discovered. For instance, Morpheus B12, D12, E12 and F12 all contain 0.1 M Tris and bicine + 37.5% mix of 2-methyl-2,4-pentanediol (MPD), PEG1000 and PEG3350 at pH 8.5, with the addition of a mix of sodiated halides (B12), ionic salts (C12), short alcohol chains (D12) or di- to penta-ethyleneglycol (E12). All of these conditions led to crystal formation, with the sodiated halides giving the best crystals. Another pattern is PACT A3-6, which all contain 0.1 M succinic acid, phosphate and glycine with pH varying from 6.0 to 9.0, and all of which leading to similar shaped crystals. A third pattern is PACT B7-9, which all contain 0.1 M MES + 20% PEG6000 with the addition of 0.2 M sodium, ammonium or lithium chloride. It is interesting to note, in this case, that the lithium chloride appears to change the shape of the crystal from rhombic to hexagonal, although this could be an artefact of the angle from which the crystal is viewed.

It can be seen that there are some common features across the three groups of crystallisation conditions. For example, Morpheus B12 contains sodiated halides, whereas PACT B7-9 contain different chloride salts, one of which is sodium chloride. Hence there is some cross over between these conditions. Likewise, Morpheus B12, D12, E12 and F12 and PACT B7-9 all contain high levels of PEG polymerised with between 1000 and 6000 monomers per molecule. Nucleation and crystal formation depends on pH and the presence of precipitants or additives as well as the protein concentration reaching saturation as vapour diffuses from the drop into the reservoir [205]. As such it can be thought that the overlap between these multiple patterns alludes to the key components that lead to the most favourable conditions for MmyJ crystallisation; namely high PEG content, mid to high pH and the presence of halide salts.

6.3.3 Diffraction Data

Once the diffractometer was operational, one of the crystals from the initial PACT B8 screen was extracted from the drop and coated in a 70:30 mixture of PACT B8 mother liquor and glycerol as a cryoprotectant [201] before being inserted into the X-ray diffractometer. As well as the size and quality of the crystal (which was over 100 µm having been growing for close to 9 months), this condition was chosen as several crystals had grown, giving the opportunity to optimise cryoprotection if insufficient in the first instance. However, the crystal survived freezing

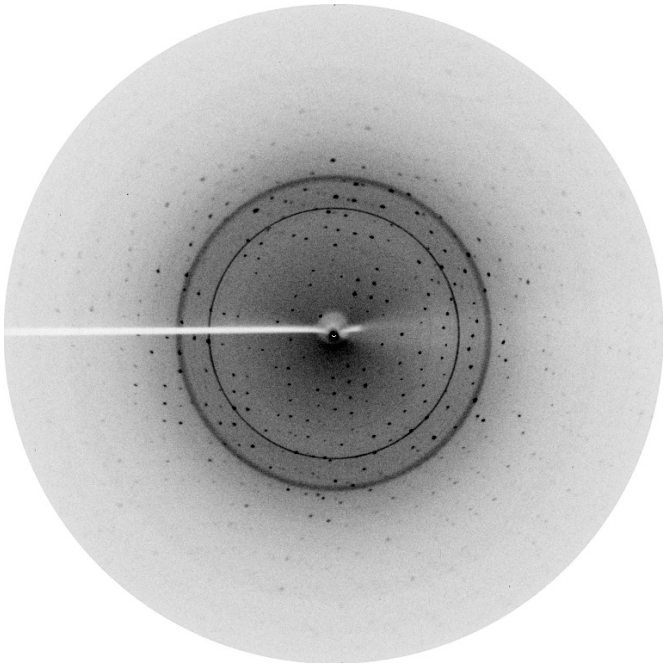


Figure 6.12: MmyJ crystal diffraction pattern with resolution of 2.1 Å, forming part of a complete set obtained over 300°.

well, and gave a good initial diffraction pattern. Repeated 15 minute exposures were then taken at 1° intervals until a coverage of 300° was obtained, giving multiple data redundancies. These data were acquired to 2.1 Å resolution, an image of which can be seen in Figure 6.12. Unfortunately, molecular replacement strategies[12] failed to solve the diffraction pattern due to a lack of protein structure in the PDB with high enough homology. The most homologous sequence as determined by BLAST analysis against the PDB was RHA00566, entry 3F6O (full name RHA1_ro06925) [118]; a suspected ArsR family protein from *Rhodococcus* species RHA1 which has had its structure solved but no functionality analyses have as yet been reported. However, this protein only has a 31% identity to MmyJ over just 54% of its sequence, and although molecular replacement was attempted, it has proved insufficient to solve the MmyJ structure.

6.3.4 Labelling Stratergies

In a further attempt to solve the diffraction data obtained, protein was expressed labelled with selenomethionine (Se-Met) in place of methionine (only occurring at M84 in MmyJ) [206]. It was hoped that this would allow the phase of the MmyJ crystal data to be determined, from

[12]Wherein attempts are made to alter the structure of a previously solved protein to match the unsolved experimental data [200].

which the previously acquired data could be solved [207, 208]. Unfortunately, Se-Met labelled MmyJ failed to form cohesive crystals under the same conditions as unlabelled MmyJ. This was despite the expectation that the same conditions should still yield high quality crystals.

Further experiments were set up to grow unlabelled crystals to be soaked in mercuric acetate in order to directly identify the crystal phase, which was the method used to solve the SmtB crystal structure [80]. However, these crystals failed to form, and only small, multifaceted crystals grew. It was thought that these could be ground up and used to seed nucleation of freshly expressed Se-Met labelled protein, but, unfortunately, there was not time to realise this experiment during the final stages of this project. As such, the obtained data remains unsolved pending the crystal phasing being identified.

7 Conclusions

7.1 Summary

7.1.1 Identification of MmyJ as an ArsR Family Transcriptional Repressor

Through bioinformatic analysis of the MmyJ amino acid sequence by BLAST, Prosite and MEME algorithms, it has been conclusively shown that MmyJ can be classified as an ArsR family transcriptional repressor. BLAST analysis identified that residues 25-83 contain an ArsR-like Helix-Turn-Helix (HTH) motif with high significance, calculating the expected value of this assignment to be 1.10×10^{-6}. By proxy, this also identified MmyJ as containing a HTH DNA binding domain, usually associated with transcription factors. It was also found that, when comparing to other amino acid sequences in the NCBI database, 80 of the 100 closest matches were ArsR family proteins, with the other 20 being hypothetical proteins of unproven function. Furthermore, of the 80 ArsR family hits, all but one had 50% identity or higher over 90 or more amino acids. This degree of similarity to proven ArsR family proteins adds significant confidence to the assignment of MmyJ as an ArsR family protein. This was further confirmed by Prosite analysis, in which it was revealed that MmyJ contains the conserved 24 residue ArsR HTH DNA binding motif, which gave good agreement when aligned against a range of 9 other ArsR family proteins. MEME analysis of the presence of this motif when compared to the same test set of 9 proven ArsR family proteins indicated that its presence in MmyJ has an expected value of 1.9×10^{-13}, granting it high significance.

Homology modelling was then undertaken using Phyre2, which predicted a structure for MmyJ containing 5 α-helices and one β-sheet. In this way, residues 49-70, corresponding to α3 and α4, along with the 3 residues between them, comprise the HTH DNA binding motif in agreement with the PROSITE assignment. This model was calculated to have high confidence throughout, with the only significant areas of disorder being at each terminus, which is not unexpected. This model was heavily based on the solution NMR structure of NmtR, a nickel and cobalt sensing transcriptional repressor of the ArsR family. HADDOCK was then used to predict the structure of the suspected MmyJ dimer based on the Phyre2 monomeric model. This dimeric model was aligned with solved crystal structures of ArsR family proteins SmtB

and RHA00566, as well as NMR structures of NmtR and CzrA, all of which aligned with an RMSD of between 4.5 and 7.4 Å over 130 atoms or more. This degree of alignment lends confidence to the predicted structure.

7.1.2 Protein Overproduction

The *mmyJ* gene was cloned into a pET151/D-TOPO expression vector for expression and overproduction of MmyJ in *E. coli*. The protein was then successfully overproduced and was found to be soluble, leading to its purification from the cell free extract (CFE) by immobilised metal affinity chromatography (IMAC) through use of the 6xHis tag incorporated by the expression vector. A tobacco etch virus (TEV) construct was obtained, with the enzyme being overproduced and purified so as to cleave the 6xHis tag from the recombinant MmyJ. It was found that the protein responsible for a band visible on an SDS-PAGE gel previously thought to be an impurity was also cleaved upon the addition of TEV. This led to the identification of a covalent dimer of MmyJ being present via the formation of a disulphide bridge between cysteines in adjacent monomers, as confirmed by assay with the reducing agent DTT. As ArsR family proteins are known to form dimers via hydrogen bonds, site directed mutagenesis was utilised to engineer a C49S mutant of MmyJ, containing no cysteines. The same DTT assay proved that there was no covalent dimer formation in this instance. However, as C49 was identified as being part of the α3-helix in the Phyre2 homology model, experiments continued with both variants of MmyJ in case the mutation of C49 caused a reduction in functionality of the protein.

7.1.3 Protein Stability Analysis

The thermal stability of MmyJ was investigated, and it was found that the secondary structure of the protein appeared to remain intact at temperatures below 40°C, with unfolding apparent at higher temperatures. The protein was also found not to refold upon cooling, as expected. However, it was noted during subsequent experiments that visible aggregates of MmyJ were forming, hence work was carried out to investigate their formation. Gel filtration chromatography identified the presence of aggregated MmyJ with a molecular weight greater than 70 kDa, equivalent to 5 or more monomers, along with a peak suspected to correspond

to the expected dimeric state. Analytical ultracentrifugation was then attempted, and a broad peak corresponding approximately to octameric oligomers of MmyJ was identified. Gel filtration chromatography was then used to investigate this oligomer formation, where it was discovered that the observed chromatagram could be replicated using boiled protein, indicating that the oligomer formation was due to unfolding of the recombinant MmyJ.

Investigations into the stability of MmyJ at high concentration found that it was stable to at least 10 mg/mL, and cryo stability experiments indicated that it was stable for a month or more under different storage conditions. However, it was found that MmyJ was extremely sensitive to multiple freeze/thaw cycles, and almost completely unfolded when frozen twice. It was thought that previous issues with oligomerisation and aggregation were the product of poor sample handling, in part due to a faulty freezer. Appropriate steps were then taken to reduce the presence of aggregates in future experiments.

7.1.4 Functional Analysis of MmyJ

Due to the genetic arrangement of *mmyJ* divergent to the *mmr* gene, itself encoding for an efflux pump that removes the antibiotic methylenomycin A (MmA) from the cell, it was suspected that MmyJ regulates expression of this gene, and hence triggers the self-resistance mechanism to MmA. Analysis of the intergenic region between *mmyJ* and *mmr* had previously revealed a 13-1-13 semi conserved inverted repeat sequence, which was found to be protected by DNA fingerprinting. Electrophoretic mobility shift assays (EMSAs) with MmyJ and fragments of the intergenic region identified this 13-1-13 motif as being the site to which MmyJ binds. As this site encompasses the -35 regions of both the *mmyJ* and *mmr* promoter regions, it can tentatively be concluded that MmyJ regulates its own expression, as well as that of *mmr*. This sequence was found to have close similarity to the typical 12-2-12 inverted repeats to which ArsR-like family members usually bind. There was no such similarity with the 12-2-12 inverted repeat to which SmtB-like family members typically bind, indicating that MmyJ is likely to bind to just a single strand of DNA, rather than both strands.

Further EMSAs were then employed in an attempt to identify the ligands sensed by MmyJ and cause the dissociation of the MmyJ:DNA complex. It was found that a mixture of MmC, D1 and D2 caused partial dissociation when present in a 20 times molar excess over MmyJ,

with the bulk of this effect found to be caused by the MmC constituent of the mixture. MmA was then found to cause total dissociation when present in the same excess. As such, it can be surmised that MmyJ regulates expression of both itself and the *mmr* gene until methylenomycin production is initiated by the cell. As methylenomycin levels increase, MmA and C are sensed by MmyJ, causing it to dissociate from the *mmr* promoter, thereby triggering production of the efflux pump and removing MmA from the cell. As methylenomycin levels then fall, new MmyJ is produced and again forms a complex with DNA at the binding site, preventing further expression and production of the MmA resistance mechanism until needed.

7.1.5 Structural Analysis of MmyJ

Secondary structural analysis of MmyJ was carried out by circular dichroism (CD) spectroscopy, and it was found that the results gave good agreement to the α-helical and β-sheet content of the homology model predicted by Phyre2. Nuclear magnetic resonance (NMR) spectroscopy was then utilised in an attempt to obtain structural information by experimental means. Although some peaks were visible in the resulting solution state spectra, it was decided that solid state NMR would be required in order to acquire data pertaining to the full, three dimensional structure of MmyJ. However, attempts at solid state NMR failed to give useable signal.

Growth conditions for the crystallisation of MmyJ were screened and hits were investigated. X-ray diffraction data was obtained to a resolution of 2.1 Å over 300°, but unfortunately these data could not be solved by molecular replacement analysis due to the absence of any suitably homologous structures in the PDB. Attempts were made to grow crystals of MmyJ labelled with selenomethione in order to determine the phase of the crystals, thus allowing the previously collected data to be solved. However, efforts in this were not successful during the timescale of the project.

7.2 Recommendations for Future Work

7.2.1 Structure Determination

Further effort should be made to grow crystal labelled with selenomethionine, utilising crushed unlabelled crystals in order to seed nucleation and crystal growth. In this way it is hoped

that the phase of the crystals can be determined, leading to the solution of the crystal data previously obtained. Once solved, this would provide conclusive structural information, to be compared to both the previously constructed homology model as well as other solved ArsR structures in order to gain deeper insights into the function of MmyJ.

It would also be desirable to continue work using NMR; either utilising more advanced pulse sequences and labelling strategies in solution state, or by growing isotopically labelled crystals for solid state. Even if the solution by X-ray diffraction is possible, complementary information from NMR, including peak assignments, would offer information into the kinetics and innate flexibility of parts of the protein.

7.2.2 DNA Binding Site

It would be advantageous to further confirm the target region of DNA to which MmyJ binds by repeating the attempted 29-mer EMSAs as detailed in Section 5.1.4, possibly with slightly longer oligonucleotides in case interactions with the β-sheets is required to stabilise the bond. This should ideally be done in a higher percentage Native-PAGE gel in order to increase retardation of the smaller DNA fragments. More work would need to be carried out to optimise the DNA:MmyJ ratio under these conditions. It is hoped that this would conclusively prove that only the single iteration of the 13-1-13 dyad is needed, without the extra half repeat shown in Figure 7.1. Also, EMSAs should be repeated using both wild type and C49S variants of MmyJ so as to investigate the significance of C49 within the α3-helix, and determine whether it improves stability of the DNA:MmyJ complex once formed.

Once the minimal length of DNA required for binding has been identified, it would be desirable to attempt co-crystallisation of MmyJ with the target DNA region in order to identify key residues required for DNA binding. This would also allow the determination of the αR-αR' distance required for DNA binding, which can be compared to that known for other ArsR family

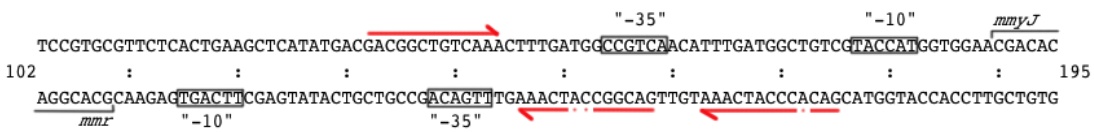


Figure 7.1: Identified DNA motif bound by MmyJ, as Figure 5.8. Red arrows indicate 13-1-13 semi conserved inverted repeat, with adjacent extra 13 bp repeat.

proteins. In this way it can be determined whether a bend in the DNA is required for complex formation, and whether MmyJ is an example of a winged HTH transcription factor, such as HlyU.

If co-crystallisation is not possible, then solution NMR titrations should be carried out, monitoring the interactions between nuclei in the protein and the ^{31}P nuclei in DNA. Assuming that a fully assigned spectrum of MmyJ can be obtained, a series of spectra taken as DNA is titrated into the NMR sample would lead to shifts in peak positions for those nuclei involved in DNA binding. In this way it would be possible to identify key residues involved in forming the complex with DNA, as well as accessing some of the binding kinetics of the interaction.

7.2.3 Ligand Identification

EMSAs attempted with MmF compounds, as well as metal ions known to bind to ArsR family proteins, should be repeated once the cause of the inconclusive positive controls has been identified and compensated for. This would provide conclusive evidence whether or not the signalling molecules that trigger methylenomycin production also trigger the resistance mechanism, or whether it is just the methylenomycins themselves that are sensed by MmyJ. Work should also continue with proven ligands MmA and C in order to determine the molar ratio needed for complete dissociation of MmyJ and DNA, from which the number of molecules of ligand required per MmyJ dimer could be determined.

Once fully tested, and all ligands have been identified, co-crystallisation and NMR titration should be attempted as described above for the interactions with DNA. In this way the ligand sensing site of MmyJ can then be determined, as well as the αR-αR' distance of the inhibitted protein, which would show how the conformation of MmyJ is changed such that DNA binding is impaired. This would allow the structural impact on the functionality of MmyJ to be fully explained. Also, co-crystallisation could lead to the conclusive identification of the method of sensing of the ligands by MmyJ; i.e. whether they physically bind to the protein or alter its structure another way, such as through a redox reaction.

7.2.4 Binding Kinetics

It was originally intended that surface plasmon resonance (SPR) would be used to determine binding kinetics of MmyJ to both DNA and binding ligand, however there has not been time to realise this. It was intended that a streptavidin SPR chip would be prepared, with biotinylated fragments of DNA immobilised on its surface. In this way, MmyJ could then be washed over the immobilised DNA, allowing the association and dissociation constants to be determined. If sufficient ligand could be produced, this could then be repeated with ligand washed over the bound MmyJ, such that the increase in dissociation upon ligand binding could be quantified. This could then be performed for all identified binding ligands, complementing the data obtained by NMR titration.

7.2.5 *In vivo* Reporter Systems

While the above *in vitro* work would lead to the full determination of the structure and function of MmyJ, the further development *in vivo* reporter systems discussed in Section 5.3.2 and 5.3.3 should not be neglected. If perfected, these systems could be utilised to engineer a novel inducible expression system for use in either *Streptomyces* or *E. coli* cultures, similar to the *lac* system used in this work. Due to its high GC content, this could provide a useful system to be used when investigating secondary metabolite or protein production in *Streptomyces* species. In this instance, MmA would be used to induce expression, requiring the inclusion of the *mmr* gene as well to provide resistance to this antibiotic.

One possible route for further developing the Tinsel Purple reporter system lies in the mutation of the *mmr* -10 site included in the intergenic region cloned into plasmid pJ251. Site directed mutation of this site as shown in Figure 7.2 would result in a stronger *E. coli* promoter, which could enable the system to be developed for use in *E. coli*. The primers required to do

Figure 7.2: Proposed mutation of *mmr* -10 site for stronger expression in *E. coli*.

this have been designed in the same way as those used for the C49S mutation, and are included as oligonucleotide pair 22 in Section 8.1.3.

7.3 Concluding Remarks

In respect to the work that was carried out, there are several comments that could be made about the approaches used. It could be considered that any one of the biophysical techniques attempted; either solution or solid-state NMR or X-ray diffraction, could have been used to elucidate a completely solved structure of MmyJ if they had been the sole focus of the work. However, by attempting to gain complementarity between the different techniques, the final result was hindered somewhat due to efforts being split between these methods. It should be noted that this is not necessarily a criticism, as it would be preferable to obtain full data sets from at least one of the NMR methods as well as X-ray diffraction in order to fully understand not just the structure of MmyJ but also its dynamic properties, and the ground work done here will make that easier in the future. Indeed, once the final step is taken to solve the crystal data obtained here, that information can then be fed back into an optimised NMR experiment (utilising more sophisticated labelling strategies, as well as more complex pulse sequences), allowing peaks to be assigned to the correct nuclei more readily. From this, it would then be possible to identify flexible regions of MmyJ and understand the motions and dynamics from which functionality of the protein is derived via the change in configuration. As such, these two techniques together would provide a powerful combined method of gaining insight into this system, and so it can be understood why both avenues were pursued.

Looking closer at some of the other aspects of the work, it can be seen that a lot of information was obtained from relatively few experiments using circular dichroism. This technique often seems underused, and it can be argued that most of the data obtained could be acquired through other means, such as NMR. However, the ease of sample preparation and short experimental time has meant that some properties of the secondary structure of MmyJ, as well as practical information about its stability, have been identified regardless of the difficulties experienced with other, more powerful techniques. In this way, it can be seen that there was merit in using this technique to obtain some of the more basic information about MmyJ, with the possibility

of further complementarity being obtained had more data had been acquired from the other approaches.

Further to this, the validity of the work carried out to create *in-vivo* reporter systems should be considered. While the processes of amplifying, digesting and ligating DNA prior to transforming and propagating cell lines, then extracting the inserted plasmids and sending for sequencing, can be thought of as a lengthy process, the skills and understanding needed for this work can be seen as fundamental to work in this field. As such, while the techniques ultimately offered no direct benefit to the project, it could be considered that by undertaking this work a greater understanding and appreciation of this corner of the scientific community has been achieved. Having said that, the ground work for these reporter systems should not be neglected, and further efforts should be made to complete this work not only to add further complementarity to any other data obtained, but to continue to work towards the secondary goal of the development of a novel inducible expression system.

It is also worth noting that the bioinformatic analyses presented here were done both thoroughly and succinctly, utilising a range of algorithms to obtain a near complete picture of how it is expected that MmyJ should look and behave. It has been mentioned at several points throughout the work that no one bioinformatic technique alone is adequate, and the complementarity strived for in the experimental work can be seen to be matched by these analyses. However, it could be said that a larger test set would improve results when comparing directly to other ArsR family proteins, specifically comparing to a greater range of family members that do not interact with metallic ligands as it was suspected, and later proved, that MmyJ interacts in some way with a non-metallic molecule. Of course, the author was limited by the data available elsewhere, and it was felt that there was not enough known about other non-metal sensing family members to form a reliable comparison between a statistically significant number of these proteins.

With reference to the aims and objectives outlined in Section 1.4, most targets have been at least partially met. It would have been satisfying to either solve the crystal structure of MmyJ or to prove ligand binding of MmA via co-crystallisation or mass spectrometry, but these results unfortunately could not be realised in the life of this project due to the allocation of time

elsewhere. The only aim not realised at all was the hope of investigating binding properties, kinetics and dynamics using a mix of NMR and SPR. However, even if MmyJ-MmA binding had been proven, there would likely not have been time to fully realise these experiments as well, and so it is perhaps the case that the initial project aims were too ambitious given the projected length of the project.

With regard to the literature review carried out in Chapter 1, this work has raised several interesting points. Firstly, as stated several times, this is the first instance of an ArsR family protein being demonstrated to sense a non-metallic ligand. It would be interesting to carry out experiments using reducing agents to investigate whether the interaction between MmyJ and MmA could be due to the epoxy ring in MmA allowing MmyJ to function as a redox switch, but this is not thought to be the case for two main reasons. Firstly, inhibition of DNA binding by MmyJ is also demonstrated in the presence of MmC, which does not have an epoxy ring nor any other functional group that is thought to easily lead to a redox interaction with MmyJ. Secondly, compared to other ArsR family proteins MmyJ is severely lacking in cysteine residues, possessing only one in helix α3. It is reported in many cases that not only is metal binding governed by cysteine residues, but so are redox sensing mechanisms such as the one demonstrated by BigR. Furthermore, if it is assumed that the homology model created is accurate, the single cysteine at residue 49 is not in a position to easily impact on the tertiary folding of MmyJ, being as it is on the outside of the α3 helix facing away from the rest of the protein. Once the crystal structure is fully elucidated, this may prove incorrect; however at present it is thought that the method of interaction between MmA and MmyJ could be based elsewhere. As such, it is felt that further experimentation as outlined in Section 7.2 would lead to the possible establishment of MmyJ as the first in a new sub-class of ArsR family proteins which sense non-metallic ligands through interactions that are not dependant on cysteine residues.

In closing, the author would like to conclude this work by stating that it is hoped that the projects are carried forward as suggested, leading to the determination of a fully resolved crystal structure as well as identifying the method of interaction (and possible binding site) of MmA with MmyJ. It is expected by the author that this will then lead to the identification of a new

class of ArsR family proteins as described above, which could lead to further understanding of the factors which trigger antibiotic resistance mechanisms, with possible lateral translation to medically relevant systems. It is also hoped that the *in vitro* reporter work can be completed, leading to the development of a novel inducible expression system for wider use within the scientific community.

8 Materials & Methods

8.1 Materials

8.1.1 Bacterial Strains

Plasmid Engineering/Storage	StarGateTop10 competent *E. coli* cells
Protein Overproduction	Invitrogen BL21 Star competent *E. coli* cells
RT-PCR Expression Assay	*Streptomyces* M145 (SCP1- SCP2-)
	Streptomyces W81 (Δ*mmfL* Δ*mmfH* Δ*mmfP*)
	Streptomyces W89 (Δ*mmyR*)
	Streptomyces W95 (Δ*mmyR* Δ*mmyD*)

8.1.2 Plasmids

pET151-*mmyJ*	Expression vector for His_6-MmyJ (amp)
pET151-*mmyJC49S*	Expression vector forHis_6-MmyJ C49S (amp)
LmJ1	Luciferase reporter system, *mmyJ-mmr* intergenic region (apr)
LmJ2	Luciferase reporter system, *mmyJ* + *mmyJ-mmr* intergenic region (apr)
pJ251	Tinsel Purple reporter system, *mmr-mmyJ* intergenic region (kan)

8.1.3 Oligonucleotide List

1. **T7 Universal Primer**
   F: TAATACGACTCACTATAGGG
2. ***mmyJ* pET151 cloning**
   F: CACCGTGGCGGCACGGATCACG
   R: TGGAGCGCCGTGCGGGCG
3. ***mmyJ* C49S Mutation**
   F: CGAAGGTGCCGCTGGCGATGTCC
   R: GAGGACATCGCCAGCGGCACCTT
4. ***mmyJ-mmr* Intergenic Region**
   F: TGCTGTTCTGCTCCTGTG
   R: CGGCTGTCCGCCTCCTCG
5. ***mmyJ* IG 110mer 1-110**
   F: TGCTGTTCTGCTCCTGTGGG
   R: CGCACGGTAATTCCGGCG

6. ***mmyJ* IG 110mer 111-218**
   F: TTCTCACTGAAGCTCATATGACG
   R: CGGCTGTCCGCCTCCTC

7. ***mmyJ* IG 50mer 101-150**
   F: TTACCGTGCGTTCTCACTGAAGCTCATATGACGACGGCTGTCAAACTTTG
   R: CAAAGTTTGACAGCCGTCGTCATATGAGCTTCAGTGAGAACGCACGGTAA

8. ***mmyJ* IG 50mer 131-180**
   F: ACGACGGCTGTCAAACTTTGATGGCCGTCAACATTTGATGGCTGTCGTAC
   R: GTACGACAGCCATCAAATGTTGACGGCCATCAAAGTTTGACAGCCGTCGT

9. ***mmyJ* IG 50mer 151-200**
   F: ATGGCCGTCAACATTTGATGGCTGTCGTACCATGGTGGAACGACACGCGA
   R: TCGCGTGTCGTTCCACCATGGTACGACAGCCATCAAATGTTGACGGCCAT

10. ***mmyJ* IG 50mer 169-218**
    F: TGGCTGTCGTACCATGGTGGAACGACACGCGACGAGGAGGCGGACAGCCG
    R: CGGCTGTCCGCCTCCTCGTCGCGTGTCGTTCCACCATGGTACGACAGCCA

11. ***mmyJ* IG 29mer 101-129 (Self Annealing)**
    GCTTACCGTGCGTTCTCACTGAAGCTCATATGCGAGGCATATGAGCTTCAG
    TGAGAACGCACGGTAAGC

12. ***mmyJ* IG 29mer 117-145 (Self Annealing)**
    GCCTGAAGCTCATATGACGACGGCTGTCAAAGCGAGGCTTTGACAGCCGTC
    GTCATATGAGCTTCAGGC

13. ***mmyJ* IG 29mer 131-159 (Self Annealing)**
    GCACGACGGCTGTCAAACTTTGATGGCCGTCGCGAGGCGACGGCCATCAAA
    GTTTGACAGCCGTCGTGC

14. ***mmyJ* IG 29mer 147-176 (Self Annealing)**
    GCTTTGATGGCCGTCAACATTTGATGGCTGTCGCGAGGCGACAGCCATCAA
    ATGTTGACGGCCATCAAAGC

15. ***mmyJ* IG 29mer 161-189 (Self Annealing)**
    GCACATTTGATGGCTGTCGTACCATGGTGGAGCGAGGCTCCACCATGGTAC
    GACAGCCATCAAATGTGC

16. ***hrdB***
    F: GATTGGGCGTAACGCTCTTG
    R: TGTCCCTGCTGGTCTTCTC

17. **SCP1.237c (*mmr*)**
    F: AGGTTGGCTCCCGTGAATC
    R: TGTCCCTGCTGGTCTTCTC

18. **SCP1.238 (*mmyJ*)**
    F: GCACGGATCACGACAGAG
    R: GTCGACGATCGCGGTGAG

19. **LmJ1 Insert (*Eco*RV → IG region → *Bam*HI)**
    F: CCGGATATCGATGCGCTCTGTCG
    R: CCGGGATCCCGCCTGTTCGGACAGTGG

20. **LmJ2 Insert (*Eco*RV → *mmyJ* + IG region → *Bam*HI)**
    F: CCGGATATCTACGGCTGTCCTCCGCAACC
    R: Same as 19R

21. **LmJ1/2 Insert Check**
    F: AAGCCACTGAGCGGGAGCTTG
    R: GACGCTGTTGTCGCCGAAGTTG

22. **pJ251 *mmr* -10 Promoter Mutation**
F: CCGGAATTACCGTGCGTTCTATTATAAGCTCATATGACGACGGCTGTCAAA CTTTG
R: CAAAGTTTGACAGCCGTCGTCATATGAGCTTATAATAGAACGCACGGTAAT TCCGG

8.1.4 Kits

Gel Extraction	Thermo Scientific GeneJET Gel Extraction Kit [209]
Plasmid Extraction	Thermo Scientific GeneJET Plasmid Miniprep Kit [210]
Reverse Transcription PCR	QIAGEN RNeasy Kit [211], Turbo DNA-free Kit [212], Superscript II Reverse TRanscriptase Kit [213], QIAquick Kit [214]
Site Directed Mutagenesis	Agilent QuikChange Lightning SDM Kit [215]

8.1.5 Gel Markers, Dyes and Stains

DNA Gel Stain	GelRed
DNA Molecular Weight Ladder	Thermo Scientific FastRuler Low Range Ladder, Thermo Scientific GeneRuler Low Range Ladder
Native PAGE Loading Dye	Expedeon DNA/Native Loading Buffer
Protein Gel Stain	Expedeon InstantBlue
Protein Molecular Weight Ladder	Spectra Multicolor Low Range Protein Ladder
SDS-PAGE Loading Dye	Expedeon LDS Loading Buffer

8.1.6 Buffer Exchange/Concentration Filters

Centrifuge Filters	Amicon Ultra-15 (10 kDa MWCO)
Dialysis	Thermo Scientific SnakeSkin Tubing (7 kDa MWCO)
Microcentrifuge Filters	Amicon Ultra 0.5 mL (10 kDa MWCO)

8.1.7 Purification Columns

Gel Filtration	Superdex 75 5/150 GL
IMAC	GE Healthcare HisTrap HP 1 mL

8.2 Instruments

Analytical Ultracentrifugation	Beckman XLI
Cell Lysis	Constant Systems Cell Disruptor (one-shot mode)
Circular Dichroism Spectroscopy	Jasco J-720
Crystallography Screen Setup	TTP Labtech Mosquito LCP
FPLC	ÄKTApurifier 10 and ÄKTA Pure 25
Mass Spectrometry	Bruker MaXis Plus
NMR (Solid State)	Bruker Avance II 600 MHz
NMR (Solution State)	Bruker Avance II 700 MHz
X-Ray Diffraction	Genix Xenocs mar 345dtpμX

8.3 Experimental Methods

8.3.1 PCR Protocols

Taq Polymerase:

- 5.0 µL 10 × Taq buffer (+ $(NH_4)_2SO_4$).
- 2.0 µL dNTPs (10 mM).
- 1.0 µL each primer.
- 0.5 µL Template.
- (3 µL $MgCl_2$ if needed).
- 2.5 µL DMSO (100%).
- 1.0 µL Taq polymerase.
- Water up to 50 µL total (37 µL if no $MgCl_2$).

1. 94°C, 2 minutes.
2. 94°C, 45 seconds.
3. 55°C, 45 seconds.
4. 72°C, 90 seconds, Go to 2. × 30.
5. 72°C, 5 minutes.
6. 4°C, hold.

QuikChange Site Directed Mutagenesis Kit:

- 5.0 µL 10× Reaction Buffer.
- 1.0 µL Template.
- 1.0 µL each primer (0.1 mM).
- 1.0 µL dNTPs (from kit).
- 3.0 µL QuikSolution.
- 38.0 µL Water.

1. 95°C, 1 minute.
2. 95°C, 50 seconds.
3. 60°C, 50 seconds.
4. 68°C, 1 minute/kb, Go to 2. × 18.
5. 68°C, 7 minutes.
6. 4°C, hold.

Roche High Fidelity Polymerase:

- 37.0 µL Water.
- 2.5 µL DMSO.
- 2.0 µL dNTPs (10 mM).
- 5.0 µL Buffer 2 (with $MgCl_2$).
- 1.0 µL Each Primer.
- 0.5 µL Template.
- 1.0 µL Polymerase.

1. 95°C, 5 minutes.
2. 95°C, 45 seconds.
3. 55°C, 45 seconds.
4. 72°C, 2 minutes, Go to 2. × 30.
5. 72°C, 15 minutes.
6. 4°C, hold.

8.3.2 Sterilisation Procedure

All work with live cells was carried out using solutions, plastics and growth media sterilised either by autoclaving or using 0.22 µm syringe filters. Cell work was carried out in a sterile environment created by first disinfecting the bench with 1% by weight Virkon solution. A bunsen burner with roaring flame was then placed in the centre of the sterilised work area in order to maintain a sterile surface.

8.3.3 Chemical Transformation of Cells

After gently thawing, 2 µL of plasmid containing solution was added to 20 µL competent cells. This mixture was kept on ice for 30 minutes, before being heat shocked at 42°C for 30 seconds, after which 500 µL of pre-warmed LB was added and the resulting cell culture was incubated at 37°C for an hour while shaking at 180 rpm. This was then split in a 1:9 ratio between two 25 mL LB plates containing 0.1 mg/mL ampicillin or other antibiotic as appropriate and left to grow at 37°C overnight.

8.3.4 Glycerol Stock Preparation

A vial of 10 mL liquid LB containing 0.1 mg/mL ampicillin or other antibiotic as appropriate was inoculated with a picked colony from a solid culture plate. This was then incubated at 37°C overnight while being shaken at 180 rpm. After incubation, 1 mL of culture was taken and added to 1 mL 50% glycerol solution for long term storage at -80°C.

8.3.5 Culture Preparation for Plasmid Extraction

A vial of 10 mL liquid LB containing 0.1 mg/mL ampicillin or other antibiotic as appropriate was inoculated either from glycerol stock or culture plate. This was then incubated at 37°C overnight while being shaken at 180 rpm. The resulting cell culture was then spun down at 4,000 rpm for 20 minutes and the supernatant discarded. The pellet was then resuspended and the plasmid extracted.

8.3.6 DNA Sequencing

20 µL of plasmid preparation was submitted according to the guidelines supplied by GATC Biotech, along with 20 µL of each primer required for **SUPREME**RUN analysis. Sequenced data was analysed using Clone Manager™ 9 Professional Edition.

8.3.7 Protein Expression

A BL21 star colony was picked and used to inoculate 10 mL LB media containing 0.1 mg/mL ampicillin, which was then grown overnight at 37°C, shaking at 180 rpm. A proportion of this starter culture was then added to a large conical flask no more than 25% full of LB (also containing 0.1 mg/mL ampicillin) such that 1% of the final liquid volume was added from the starter culture (i.e. if growing a 800 mL culture in a 5 L flask, 8 mL of the starter culture was added). This was then left in an incubator at 37°C shaking at 180 rpm until an optical density at 600 nm (OD_{600}) of about 0.6 was reached, after which time protein overproduction was initiated using IPTG at a final concentration of 1 mM. This was then incubated overnight at 15°C, shaking at 180 rpm to allow the BL21 star ample time to overproduce the protein.

The following morning, the cell culture was split into 500 mL flasks and spun down in a Sorvall RC6 Plus ultra-centrifuge at 5,000 rpm for 20 minutes. The resulting cell pellet(s) were then resuspended in IMAC wash/binding buffer (see Section 8.3.9) such that the total volume of all resupsended cells from a single culture was approximately 50 mL. This was placed in a 50 mL Falcon Tube and refrigerated until purification.

If expressing isotopically labelled proteins, the above procedure was modified as follows, using M9 minimal media. Once the cell cultures had reached OD_{600} of approximately 0.6, they were spun down at 5,000 rpm for 20 minutes, before being washed by resuspending in approximately 100 mL M9 salts. They were then centrifuged again and were resupsended in M9 media, using half the volume of the original LB culture. This was done to minimise the cost of labelled carbon and nitrogen sources used, as it has been shown that when expressing labelled proteins using M9 the yield of expressed protein is similar whether the cells are transferred into the same volume as they had been grown in LB or half that volume [216]. These cultures were then induced with IPTG and from this point onwards were handled the same as natural

abundance expression in LB.

If expressing proteins labelled with specific amino acids, such as selenomethionine in place of methionine, the above protocol for expressing isotopically labelled proteins was followed with a few minor alterations. Firstly, natural abundance carbon and nitrogen sources were used, so no isotopic labelling was introduced. Secondly, M9 media was used containing 20 mg/L of each amino acid individually, with the necessary substitution made. Finally, if required 1 mM tris(2-carboxyethyl)phosphine (TCEP) was added to all media and buffers as a reducing agent. In the case of selenomethionine labelling, this prevented aggregation via interactions between selenium in adjacent molecules.

Long Term Storage Buffer:

- 100 mM NaCl
- 20 mM TRIS (pH 8)
- 30% Glycerol

Phosphate Buffer:

- Makes 0.1 M Potassium Phosphate Buffer at desired pH using the following volume (mL) of 1 M stocks and diluted to 1 L:

pH	K_2HPO_4	KH_2PO_4	pH	K_2HPO_4	KH_2PO_4
5.8	8.5	91.5	7.0	61.5	38.5
6.0	13.2	86.8	7.2	71.1	28.3
6.2	19.2	80.8	7.4	80.2	19.8
6.4	27.8	72.2	7.6	86.6	13.4
6.6	38.1	61.9	7.8	90.8	9.2
6.8	49.7	50.3	8.0	94.0	6.0

8.3.8 Cell Lysis

Once resuspended, cells were lysed at approximately 20 kPsi using a cell disruptor. Once the entire sample had been lysed, the cell debris was then pelleted by spinning down in a Sorval RC6 Plus ultra-centrifuge at 18,000 rpm for 20 minutes. The supernatant was then passed through a 0.22 µm syringe filter to remove any remnants of cell debris or unlysed cells. This cell free extract (CFE) was then either further purified or refrigerated until needed.

8.3.9 Protein Purification

A His-Trap column was washed and allowed to equilibrate with 10 mL of wash/binding buffer as per the column's instruction booklet. A Superloop™ system was then used to pass up to 50 mL of CFE through the column, loading it with His_6-MmyJ. After this, more wash/binding

buffer was passed through the column to remove any last unbound proteins and return the UV absorbance trace to close to the baseline. The His_6-MmyJ was then eluted from the column by gradual introduction of elution buffer containing 200 mM imidazole. 3 mL fractions were collected throughout the purification process.

Ni^{2+} IMAC Wash/Binding Buffer:

- 100 mM NaCl
- 20 mM TRIS (pH 8)
- 0-10 mM Imidazole
- 10% Glycerol
- Degas before use

Ni^{2+} IMAC Elution Buffer:

- 100 mM NaCl
- 20 mM TRIS (pH 8)
- 200 mM Imidazole
- 10% Glycerol
- Degas before use

8.3.10 Protein Visualisation via SDS-PAGE

Gels were cast using either a 10 or 15 well comb depending on requirements, and were pre-run for up to 2 hours at 200 V to ensure all remaining monomeric acrylamide had been drawn out of the gel. 15 µL of each sample was then loaded and the gel was run at 200 V until the loading dye could be seen to be nearing the bottom of the gel. Gels were then stained for between 15 minutes and an hour before being de-stained in cold water overnight.

SDS-PAGE Gels:

(Makes 60 mL)	**15%**	**12%**	**4%**	**4% (2 mL)**
1.5 M TRIS (pH 8.8)	15.0 mL	15.0 mL	15.0 mL	0.5 mL
20% SDS	0.3 mL	0.3 mL	0.3 mL	30 µL
10% Acrylamide	30.0 mL	24.0 mL	7.5 mL	0.25 mL
Water	14.7 mL	20.7 mL	37.2 mL	1.25 mL

- For 5 mL gel, add 100 µL APS and 5 µL TEMED

SDS-PAGE Running Buffer (10 ×):

- 250 mM TRIS-HCl pH 8.8
- 2 M Glycine
- 1 % SDS

8.3.11 Cleavage by TEV Protease

TEV protease was added to His_6-MmyJ at an approximately 1:100 TEV:His_6-MmyJ molar ratio, which was then incubated at room temperature overnight on a rocking plate oscillating

approximately once every 10 seconds. This mixture was then purified via FPLC similar to CFE purification. However in this case the TEV protease was retained by the HisTrap column, along with the cleaved 6xHis tags and other protein impurities. This allowed the collection of pure, cleaved MmyJ as a non-specific binding protein by elution with 20 mM imidazole.

TEV Buffer:

- 50 mM TRIS (pH 8).
- 0.5 mM EDTA.
- 1 mM DTT.

8.3.12 Site Directed Mutagenesis

Primer pair 3 were designed using the online QuikChange Primer Design tool [217]. PCR was carried out using the protocol 8.3.1 and XL10-Gold competent cells supplied with the QuikChange Kit were transformed with the PCR product according to [215] and solid LB plates containing 0.1 mg/mL ampicillin were inoculated with the transformed cells as usual.

8.3.13 Mass Spectrometry Sample Preparation and Analysis

Protein sample was buffer exchanged into 50 mM ammonium acetate and diluted to a final concentration of approximately 1 µM . This was then loaded by direct infusion. Data was analysed using Bruker DataAnalysis, and spectra were deconvoluted using the maximum entropy algorithm.

8.3.14 Buffer Exchange by Dialysis

Protein solution was put into an appropriate length of dialysis tubing. This was then sealed and submerged in 2 L of new buffer and left for at least 6 hours, before the buffer was replaced and it was left for another 6 hours. In cases where more than 10 mL of protein solution was to be dialysed, a 5 L reservoir of new buffer was used, such that the ratio of new buffer to old never dropped below 200:1.

8.3.15 Circular Dichroism Spectroscopy

Protein samples were prepared to a concentration of 0.5 mg/mL in phosphate buffer pH6.5 unless otherwise stated. In order to minimise the amount of sample needed, a 1 mm quartz

cuvette was used, and 3 acquisitions were run in order to reduce noise. As N_2 used to purge the instrument and H_2O in the sample buffer both absorb strongly at less than 180 nm [146] and natural amino acids do not give rise to CD bands at wavelengths above 300 nm [144], only wavelengths between these values were scanned. Unless otherwise stated, all spectra were taken at 25°C.

8.3.16 Gel Filtration Chromatography

A column was used with an optimum protein separation range of 3 kDa to 70 kDa and a 50 µL maximum sample volume. As each His_6-MmyJ monomer has a mass of approximately 15 kDa, this mass range was thought to be suitable for investigating the protein oligomeric state. Also, the small sample volume meant that analytical work could be carried out quickly without using up large quantities of protein, as would be the case in a larger column which would be more appropriate for purification. The UV absorption at 280 nm was recorded as a function of cumulative volume passed through the column.

A series of solutions containing proteins of known mass was run in order to calibrate the column in the configuration used. These proteins were from the Gel Filtration Molecular Weight Markers Kit for Molecular Weights 6,500-66,000 Da produced by Sigma-Aldrich ®, which were dissolved in gel filtration buffer as per the kit's instructions. These solutions were then run through the FPLC machines as 50 µL samples in order to calibrate the column (see Figure 4.4, Section 4.2.1).

Gel Filtration Buffer:

- 100 mM NaCl
- 20 mM TRIS (pH 8)
- 10% Glycerol
- Degas before use

8.3.17 Analytical Ultracentrifugation

Protein samples were prepared in 50 mM TRIS pH 8.8 + 0.5 mM EDTA with 10% glycerol by weight added for stability. This was then sent to the Birmingham Biophysical Characterisation Facility (BBCF) to be run. The data from this was then analysed using Sedfit [218] alongside theoretical parameters for buffer density (ρ) and viscosity (η) calculated using Sednterp [219],

as well as partial specific volume ($\bar{V}$) calculated from the amino acid sequence.

8.3.18 Agarose Gels

Agarose was weighed out to either 1.0 g/L or 1.2 g/L depending on size of DNA fragment and TBE buffer was added as required. Mixture was then heated in a microwave at 450 W until melted, after which it was held under cold running water for approximately a minute. This cooled the gel enough to add DNA intercalating stain at 5 µL per 100 mL of gel. Gel was then poured into the casting tray with an appropriate comb and allowed to set.

TBE Buffer (5 ×):

- 27 g TRIS.
- 14 g Boric Acid.
- 3.4 g EDTA.
- (HCl to pH 8-8.3.)
- Water to 0.5 L.

8.3.19 Electrophoretic Mobility Shift Assay (No Ligand)

Unless otherwise stated, 1 µL DNA was incubated with 10 µL MmyJ in IMAC wash/binding buffer (or 10 µL of buffer for negative controls) plus 1 µL EMSA buffer and 8 µL dH_2O for 15 minutes at room temperature. After this time loading dye was added and the assay was loaded into a Native PAGE gel and run at 100 V until the stain neared the bottom of the gel. NB: If using annealed oligonucleotides, DNA was first diluted by a factor of 10

Native PAGE Gel (6%):

- 2 mL 5× TBE Buffer
- 2 mL 30% Acrylamide
- 6 mL dH_2O
- 200 µL APS
- 10 µL TEMED
- Pre-run gel for 2 hours at 120 V before use.

8.3.20 Electrophoretic Mobility Shift Assay (Ligand Added)

As above, but after the 15 minute incubation 1 µL of suspected ligand was added (or 1 µL solvent for negative control) and the mixture was then left to incubate for a further 15 minutes. Loading dye was then added and the gel was loaded and run as above.

8.3.21 Reverse Transcription PCR: cDNA Production

Streptomyces strains were grown in SMM media for 3 days at 30°C while undergoing shaking at 200 rpm. These cells were then pelleted, resuspended and lysed on a vortex using tubes containing Lysing Matrix E from the FastDNA™ SPIN Kit for Soil by MP Biomedicals [220]. cDNA was then purified using the QIAGEN RNeasy Kit [211], Turbo DNA-free Kit [212] and SuperScript II Reverse Transcriptase [213], before using the QIAquick Kit [214] to clean up the resulting cDNA. This was then used as a template for PCR under normal conditions.

8.3.22 *Bam*HI and *Eco*RV Double Digestion

1 µL each of NEB *Bam*HI and *Eco*RV enzyme was added to 16 µL DNA along with 2 µL CutSmart buffer. Alkaline phosphatase was added to the reaction mixture to prevent religation where required. The resulting mixture was then incubated at 37°C for 30 minutes. Digested DNA was then gel purified as needed.

8.3.23 Plasmid Ligation

Ligase was used in accordance with the protocol supplied with the enzyme. Resulting ligated plasmid was transformed into Top10 cells and plated onto LB media containing 0.05 mg/mL apramycin and incubated at 37°C overnight

8.4 Growth Media

LB Media:

- 10.0 g Bactatryptone.
- 5.0 g Yeast extract.
- 10.0 g NaCl.
- (15.0 g Agar for solid medium).
- 0.1% Antibiotic (add after autoclaving).
- Water up to 1.0 L.

M9 Salt Solution (5 ×):

- 32.0 g $Na_2HPO_4 \cdot 7H_2O$, or 17.0 g Na_2HPO_4
- 6.0 g KH_2PO_4
- 2.5 g NaCl
- 5.0 g NH_4 salt, unless adding label, in which case add after mixed
- Water up to 0.5 L

M9 Media:

- 100 mL 5 × M9 salt solution.
- 1 mL 1 M $MgSO_4$ solution.
- 50 µL 1 M $CaCl_2$ solution.
- 0.5 g Nitrogen source (add after autoclaving).
- 1 g Carbon source (add after autoclaving).
- 1% 100× BME vitamins (add after autoclaving).
- Tap Water up to 0.5 L.

Soya Flour Media:

- 10.0g Mannitol.
- 10.0 g Soya flour.
- 10.0 mM $CaCl_2$.
- (10.0 g Agar for solid medium).
- Water up to 0.5 L.

9 References

[1] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, *Molecular Biology of the Cell*. Garland Publishing Inc, 1989.

[2] D. S. Latchman, "Transcription Factors: An Overview," *Int. J. Biochem. Cell B.*, vol. 29, no. 12, pp. 1305–12, 1997.

[3] R. G. Roeder, "The Role of General Initiation Factors in Transcription by RNApolymerase II," *Trends Biochem. Sci.*, vol. 21, no. 9, pp. 327–35, 1997.

[4] D. S. Latchman, *Eukaryotic Transcription Factors*, 3rd ed. Academic Press, 1998.

[5] L. Escolar, J. Perez-Martin, and V. de Lorenzo, "Opening the Iron Box: Transcriptional Metalloregulation by the Fur Protein," *J. Bacteriol.*, vol. 181, no. 20, pp. 6223–9, 1999.

[6] B. P. Rosen, U. Weigel, C. Karkaria, and P. Gangola, "Molecular Characterisation of an Anion Pump," *J. Biol. Chem.*, vol. 263, no. 7, pp. 3067–70, 1988.

[7] J. Wu and B. P. Rosen, "The ArsR Protein is a Trans-Acting Regulatory Protein," *Mol. Microbiol.*, vol. 5, no. 6, pp. 1331–6, 1991.

[8] A. L. Taylor and C. D. Trotter, "Linkage Map of Escherichia coli Strain K-12," *Bacteriol. Rev.*, vol. 36, no. 4, pp. 504–24, 1972.

[9] A. Jobe and S. Bourgeois, "Lac Repressor-Operator Interaction: VI. The Natural Inducer of the lac Operon," *J. Mol. Biol.*, vol. 69, no. 3, pp. 397–408, 1972.

[10] W. Gilbert and B. Muller-Hill, "Isolation of the lac Repressor," *Proc. Nat. Acad. Sci. USE*, vol. 56, no. 6, pp. 1891–8, 1966.

[11] A. Marbach and K. Bettenbrock, "lac Operon Induction in Escherichia coli: Systematic Comparison of IPTG and TMG Induction and Influence of the Transacetylase LacA," *J. Biotechnol.*, vol. 157, pp. 82–8, 2012.

[12] F. Studier, A. Rosenburg, J. Dunn, and J. Dubendorff, "Use of T7 RNA Polymerase to Direct Expression of Cloned Genes," *Meth. Enzymol.*, vol. 185, pp. 60–89, 1990.

[13] J. Dubendorff and F. Studier, "Controlling Basal Expression in an Inducible T7 Expression System by Blocking the Target T7 Promoter with lac Repressor," *J. Mol. Biol.*, vol. 219, pp. 45–59, 1991.

[14] D. B. McKay and T. A. Steitz, "Structure of Catabolite Gene Activator Protein at 2.9 A Resolution Suggests Binding to Left-Handed B-DNA," *Nature*, vol. 290, no. 5809, pp. 744–9, 1981.

[15] W. F. Anderson, D. H. Ohlendorf, Y. Takeda, and B. W. Matthews, "Structure of the cro Repressor from Bacteriophage Lambda and its Interaction with DNA," *Nature*, vol. 290, no. 5809, pp. 754–8, 1981.

[16] B. W. Matthews, D. H. Ohlendorf, W. F. Anderson, and Y. Takeda, "Structure of the DNA-Binding Region of lac Repressor Inferred from its Homology with cro Repressor," *Proc. Nat. Acad. Sci. USE*, vol. 79, no. 5, pp. 1428–32, 1982.

[17] C. O. Pabo and M. Lewis, "The Operator-Binding Domain of Lambda Repressor: Structure and DNA Recognition," *Nature*, vol. 298, no. 5873, pp. 443–7, 1982.

[18] C. O. Pabo and R. T. Sauer, "Protein-DNA Recognition," *Ann. Rev. Biochem.*, vol. 53, pp. 293–321, 1984.

[19] S. C. Harrison and A. K. Aggarwal, "DNA Recognition by Proteins with the Helix-Turn-Helix Motif," *Ann. Rev. Biochem.*, vol. 59, pp. 933–69, 1990.

[20] R. Wintjens and M. Rooman, "Structural Classification of HTH DNA-binding Domains and Protein–DNA Interaction Modes," *J. Mol. Biol.*, vol. 262, pp. 294–313, 1996.

[21] T. A. Steitz, D. H. Ohlendorf, D. B. McKay, W. F. Anderson, and B. W. Matthews, "Structural Similarity in the DNA-Binding Domains of Catabolite Gene Activator and cro Repressor Proteins," *Proc. Nat. Acad. Sci. USE*, vol. 79, no. 10, pp. 3097–3100, 1982.

[22] R. G. Brennan and B. W. Matthews, "The Helix-Turn-Helix DNA Binding Motif," *J. Biol. Chem.*, vol. 264, no. 4, pp. 1903–6, 1989.

[23] A. Marchler-Bauer, M. K. Derbyshire, N. R. Gonzales, S. Lu, F. Chitsaz, L. Y. Geer, R. C. Geer, J. He, M. Gwadz, D. I. Hurwitz, C. J. Lanczycki, F. Lu, G. H. Marchler, J. S. Song, N. Thanki, Z. Wang, R. A. Yamashita, D. Zhang, C. Zheng, and S. H. Bryant, "CDD: NCBI's Conserved Domain Database," *Nucleic Acids Res.*, vol. 43, pp. 222–6, 2015.

[24] NCBI CDD Conserved Protein Domain HTH. [Online]. Available: http://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=277524

[25] J. L. Ramos, M. Martinez-Bueno, A. J. Molina-Henares, W. Teran, K. Wantabe, X. Zhang, M. Gallegos, R. G. Brennan, and R. Tobes, "The TetR Family of Transcriptional Repressors," *Microbiol. Mol. Biol. Rev.*, vol. 69, no. 2, pp. 326–56, 2005.

[26] L. Cuthbertson and J. R. Nodwell, "The TetR Family of Regulators," *Microbiol. Mol. Biol. Rev.*, vol. 77, no. 3, pp. 440–75, 2013.

[27] R. G. Martin and J. L. Rosner, "The AraC Transcriptional Activators," *Curr. Opin. Microbiol.*, vol. 4, pp. 132–7, 2001.

[28] L. S. Busenlehner, M. A. Pennella, and D. P. Giedroc, "The SmtB/ArsR Family of Metalloregulatory Transcriptional Repressors: Structural Insights into Prokaryotic Metal Resistance," *FEMS Microbiol. Rev.*, vol. 27, pp. 131–43, 2003.

[29] K. Yokoyama, S. A. Ishijima, L. Clowney, H. Koike, H. Aramaki, C. Tanaka, K. Makino, and M. Suzuki, "Feast/Famine Regulatory Proteins (FFRPs): Escherichia coli Lrp, AsnC and Related Archaeal Transcription Factors," *FEMS Microbiol. Rev.*, vol. 30, pp. 89–108, 2006.

[30] H. Korner, H. J. Sofia, and W. G. Zumft, "Phylogeny of the Bacterial Superfamily of Crp-Fnr Transcription Regulators: Exploiting the Metabolic Spectrum by Controlling Alternative Gene Programs," *FEMS Microbiol. Rev.*, vol. 27, pp. 559–92, 2003.

[31] G. Zeng, S. Ye, and T. J. Larson, "Repressor for the Sn-Glycerol 3-Phosphate Regulon of Escherichia coli K-12: Primary Structure and Identification of the DNA-Binding Domain," *J. Bacteriol.*, vol. 178, pp. 7080–9, 1996.

[32] S. Rigali, A. Derouaux, F. Giannotta, and J. Dusart, "Subdivision of the Helix-Turn-Helix GntR Family of Bacterial Regulators in the FadR, HutC, MocR, and YtrA Subfamilies," *J. Biol. Chem.*, vol. 277, pp. 12 507–15, 2002.

[33] A. J. Molina-Henares, "Members of the IclR Family of Bacterial Transcriptional Regulators Function as Activators and/or Repressors," *FEMS Microbiol. Rev.*, vol. 30, pp. 157–86, 2006.

[34] L. Swint-Kruse and K. S. Matthews, "Allostery in the LacI/GalR Family: Variations on a Theme," *Curr. Opin. Microbiol.*, vol. 12, pp. 129–37, 2009.

[35] J. Chen and J. Xie, "Role and Regulation of Bacterial LuxR-Like Regulators," *J. Cell. Biochem.*, vol. 112, pp. 2694–702, 2011.

[36] S. E. Maddocks and P. C. Oyston, "Structure and Function of the LysR- Type Transcriptional Regulator (LTTR) Family Proteins," *Microbiology*, vol. 154, pp. 3609–23, 2008.

[37] S. P. Wilkinson and A. Grove, "Ligand-Responsive Transcriptional Regulation by Members of the MarR Family of Winged Helix Proteins," *Curr. Issues Mol. Biol.*, vol. 8, pp. 51–62, 2006.

[38] J. L. Hobman, J. Wilkie, and N. L. Brown, "A Design for Life: Prokaryotic Metal-Binding MerR Family Regulators," *Biometals*, vol. 18, pp. 429–36, 2005.

[39] S. Kustu, A. K. North, and D. S. Weiss, "Prokaryotic Transcriptional Enhancers and Enhancer-Binding Proteins," *Trends Biochem. Sci.*, vol. 16, pp. 397–402, 1991.

[40] M. Martinez-Bueno, R. Tobes, and J. L. Ramos, "Structural Relationships in the OmpR Family of Winged-Helix Transcription Factors," *J. Mol. Biol.*, vol. 269, pp. 301–12, 1997.

[41] L. J. Shimon and S. C. Harrison, "The phage 434 OR2/R1-69 Complex at 2.5 A Resolution," *J. Mol. Biol.*, vol. 232, pp. 826–38, 1993.

[42] L. L. C. Schrodinger, "The PyMOL Molecular Graphics System, Version 1.7.4 Schrödinger, LLC."

[43] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, "The Protein Data Bank," *Nucleic Acids Res.*, vol. 28, pp. 235–42, 2000.

[44] T. B. K. Le, C. E. M. Stevenson, H.-P. Fiedler, A. Maxwell, D. M. Lawson, and M. J. Buttner, "Structures of the TetR-like Simocyclinone Efflux Pump Repressor, SimR, and the Mechanism of Ligand-Mediated Derepression," *J. Mol. Biol.*, vol. 408, pp. 40–56, 2011.

[45] T. B. K. Le, M. A. Schumacher, D. M. Lawson, R. G. Brennan, and M. J. Buttner, "The Crystal Structure of the TetR Family Transcriptional Repressor SimR Bound to DNA and the Role of a Flexible N-terminal Extension in Minor Groove Binding," *Nucleic Acids Res.*, vol. 39, no. 21, pp. 9433–47, 2011.

[46] M. J. D. SanFrancisco, C. L. Hope, J. B. Owolabi, L. S. Tisa, and B. P. Rosen, "Identification of the Metalloregulatory Element of the Plasmid-Encoded Arsenical Resistance Operon," *Nucleic Acids Res.*, vol. 18, no. 3, pp. 619–24, 1990.

[47] S. Silver, G. Ji, S. Broer, S. Dey, D. Dou, and B. P. Rosen, "Orphan Enzyme or Patriarch of a New Tribe: The Arsenic Resistance ATPase of Bacterial Plasmids," *Mol. Microbiol.*, vol. 8, no. 4, pp. 637–42, 1993.

[48] G. Nucifora, L. Chu, T. K. Misra, and S. Silver, "Cadmium Resistance of Staphylococcus aureus Plasmid p1258 Results from a Cd2+ Efflux ATPase Determined by the cadA Gene," *Proc. Nat. Acad. Sci. USE*, vol. 86, pp. 3544–8, 1989.

[49] S. Silver and M. Walderhaug, "Gene Regulation of Plasmid- and Chromosome-Determined Inorganic Ion Transport in Bacteria," *Microbiol. Rev.*, vol. 56, no. 1, pp. 195–228, 1992.

[50] D. R. Harvie, C. Andreini, G. Cavallaro, W. Meng, B. A. Connolly, K. ichi Yoshida, Y. Fujita, C. R. Harwood, D. S. Radford, S. Tottey, J. S. Cavet, and N. J. Robinson, "Predicting Metals Sensed by ArsR-SmtB Repressors: Allosteric Interference by a Non-Effector Metal," *Mol. Microbiol.*, vol. 59, no. 4, pp. 1341–56, 2006.

[51] D. Osman and J. S. Cavet, "Bacterial Metal-Sensing Proteins Exemplified by ArsR-SmtB Family Repressors," *Nat. Prod. Rep.*, vol. 27, no. 5, pp. 668–80, 2010.

[52] A. P. Morby, J. S. Turner, J. W. Huckle, and N. J. Robinson, "SmtB is a Metal-Dependent Repressor of the Cyanobacterial Metallothionein Gene smtA: Identification of a Zn Inhibited DNA-Protein Complex," *Nucleic Acids Res.*, vol. 21, no. 4, pp. 921–5, 1993.

[53] J. W. Huckle, A. P. Morby, J. S. Turner, and N. J. Robinson, "Isolation of a Prokaryotic Metallothionein Locus and Analysis of Transcriptional Control by Trace Metal Ions," *Mol. Microbiol.*, vol. 7, no. 2, pp. 177–87, 1993.

[54] W. Shi, J. Wu, and B. P. Rosen, "Identification of a Putative Metal Binding Site in a New Family of Metalloregulatory Protein," *J. Biol. Chem.*, vol. 269, no. 31, pp. 19 826–9, 1994.

[55] A. Bairoch, "A Possible Mechanism for Metal-Ion Induced DNA-Protein Dissociation in a Family of Prokaryotic Transcriptional Regulators," *Nucleic Acids Res.*, vol. 21, no. 10, p. 2515, 1993.

[56] R. Finn, A. Bateman, J. Clements, P. Coggill, R. Eberhardt, S. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. Sonnhammer, J. Tate, and M. Punta, "The Pfam Protein Families Database," *Nucleic Acids Res.*, vol. 42, no. Database Issue, pp. D222–30, 2014.

[57] D. R. Campbel, K. E. Chapman, K. J. Waldron, S. Tottey, S. Kendall, G. Cavallaro, C. Andreini, J. Hinds, N. G. Stoker, N. J. Robinson, and J. S. Cavet, "Mycobacterial Cells Have Dual Nickel-Cobalt Sensors," *J. Biol. Chem.*, vol. 282, no. 44, pp. 32 298–310, 2007.

[58] S. Mandal, S. Chatterjee, B. Dam, P. Roy, and S. K. DasGupta, "The Dimeric Repressor SoxR Binds Cooperatively to the Promoter(s) Regulating Expression of the Sulfur Oxidation (sox) Operon of Pseudaminobacter salicylatoxidans KCT001," *Microbiology*, vol. 153, pp. 80–91, 2007.

[59] J. Wu and B. P. Rosen, "Metalloregulated Expression of the ars Operon," *J. Biol. Chem.*, vol. 268, no. 1, pp. 52–8, 1993.

[60] C. Xu, W. Shi, and B. P. Rosen, "The Chromosomal arsR Gene of Escherichia coli Encodes a trans-acting Metalloregulatory Protein," *J. Biol. Chem.*, vol. 271, no. 5, pp. 2427–34, 1996.

[61] C. Xu and B. P. Rosen, "Dimerization is Essential for DNA Binding and Repression by the ArsR Metalloregulatory Protein of Escherichia coli," *J. Biol. Chem.*, vol. 272, no. 25, pp. 15 734–8, 1997.

[62] J. D. Gralla, "Promoter Recognition and mRNA Initiation by Escherichia coli Eo70," *Meth. Enzymol.*, vol. 185, pp. 37–54, 1990.

[63] B. Gicquel-Sanzey and P. Cossart, "Homologies Between Different Procaryotic DNA-Binding Regulatory Proteins and Between their Sites of Action," *EMBO J.*, vol. 1, no. 5, pp. 591–5, 1982.

[64] M. A. E. Churchill and A. A. Travers, "Protein Motifs that Recognise Structural Features of DNA," *Trends Biochem. Sci.*, vol. 16, pp. 92–7, 1991.

[65] R. Rosenstein, K. Nikoleit, and F. Götz, "Binding of ArsR, the Repressor of the Staphylococcus xylosus (pSX267) Arsenic Resistance Operon to a Sequence with Dyad Symmetry Within the ars Promoter," *Mol. Gen. Genet.*, vol. 242, no. 5, pp. 566–72, 1994.

[66] J. L. Erbe, K. B. Taylor, and L. M. Hai, "Metalloregulation of the Cyanobacterial smt Locus: Identification of SmtB Binding Sites and Direct Interaction with Metals," *Nucleic Acids Res.*, vol. 23, no. 13, pp. 2472–8, 1995.

[67] S. R. Kar, A. C. Adams, J. Lebowitz, K. B. Taylor, and L. M. Hall, "The Cyanobacterial Repressor SmtB Is Predominantly a Dimer and Binds Two Zn2+ Ions per Subunit," *Biochemistry*, vol. 36, pp. 15 343–8, 1997.

[68] V. K. Singh, A. Xiong, T. R. Usgaard, S. Chakrabarti, R. Deora, T. K. Misra, and R. K. Jayaswal, "ZntR is an Autoregulatory Protein and Negatively Regulates the Chromosomal Zinc Resistance Operon znt of Staphylococcus aureus," *Mol. Microbiol.*, vol. 33, no. 1, pp. 200–7, 1999.

[69] L. S. Busenlehner, N. J. Cosper, R. A. Scott, B. P. Rosen, M. D. Wong, and D. P. Giedroc, "Spectroscopic Properties of the Metalloregulatory Cd(II) and Pb(II) Sites of S. aureus pI258 CadC," *Biochemistry*, vol. 40, pp. 4426–36, 2001.

[70] L. S. Busenlehner, T.-C. Weng, J. E. Penner-Hahn, and D. P. Giedroc, "Elucidation of Primary (a3N) and Vestigial (a5) Heavy Metal-binding Sites in Staphylococcus aureus pI258 CadC: Evolutionary Implications for Metal Ion Selectivity of ArsR/SmtB Metal Sensor Proteins," *J. Mol. Biol.*, vol. 319, no. 3, pp. 685–701, 2002.

[71] M. Bose, D. Slick, M. J. Sarto, P. Murphy, D. Roberts, J. Roberts, and R. D. Barber, "Identification of SmtB/ArsR cis Elements and Proteins in Archaea Using the Prokaryotic InterGenic Exploration Database (PIGED)," *Archaea*, vol. 2, no. 39-49, 2006.

[72] T. Liu, X. Chen, Z. Ma, J. Shokes, L. Hemmingsen, R. A. Scott, and D. P. Giedroc, "A CuI-Sensing ArsR Family Metal Sensor Protein with a Relaxed Metal Selectivity Profile," *Biochemistry*, vol. 47, no. 40, pp. 10 564–75, 2008.

[73] J. S. Cavet, A. I. Graham, W. Meng, and N. J. Robinson, "A Cadmium-Lead-sensing ArsR-SmtB Repressor with Novel Sensory Sites," *J. Biol. Chem.*, vol. 278, no. 45, pp. 44 560–6, 2003.

[74] A. Xiong and R. K. Jayaswal, "Molecular Characterization of a Chromosomal Determinant Conferring Resistance to Zinc and Cobalt Ions in Staphylococcus aureus," *J. Bacteriol.*, vol. 180, no. 16, pp. 4024–9, 1998.

[75] J. S. Cavet, W. Meng, M. A. Pennella, R. J. Appelhoff, D. P. Giedroc, , and N. J. Robinson, "A Nickel-Cobalt-sensing ArsR-SmtB Family Repressor," *J. Biol. Chem.*, vol. 277, no. 41, pp. 38 441–8, 2002.

[76] T. Liu, J. W. Golden, and D. P. Giedroc, "A Zinc(II)/Lead(II)/Cadmium(II)-Inducible Operon from the Cyanobacterium anabaena Is Regulated by AztR, an a3N ArsR/SmtB Metalloregulator," *Biochemistry*, vol. 44, no. 24, pp. 8673–83, 2005.

[77] J. Ye, A. Kandegedara, P. Martin, and B. P. Rosen, "Crystal Structure of the Staphylococcus aureus pI258 CadC Cd(II)/Pb(II)/Zn(II)-Responsive Repressor," *J. Bacteriol.*, vol. 187, no. 12, pp. 4214–21, 2005.

[78] C. Thelwell, N. J. Robinson, and J. S. Turner-Cavet, "An SmtB-like Repressor from Synechocystis PCC 6803 Regulates a Zinc Exporter," *Proc. Nat. Acad. Sci. USE*, vol. 95, pp. 10 728–33, 1998.

[79] T. Frickey and A. N. Lupas, "CLANS: A Java Application for Visualising Protein Families Based On Pairwise Similarity," *Bioinformatics*, vol. 20, pp. 3702–4, 2004.

[80] W. J. Cook, S. R. Kar, K. B. Taylor, and L. M. Hall, "Crystal Structure of the Cyanobacterial Metallothionein Repressor SmtB: a Model for Metalloregulatory Proteins," *J. Mol. Biol.*, vol. 275, no. 2, pp. 337–46, 1998.

[81] A. I. Arunkumar, G. C. Campanello, and D. P. Giedroc, "Solution Structure of a Paradigm ArsR Family Zinc Sensor in the DNA-Bound State," *Proc. Nat. Acad. Sci. USE*, vol. 106, no. 43, pp. 18 177–82, 2009.

[82] R. G. Brennan, "The Winged-Helix DNA-Binding Motif: Another Helix-Turn-Helix Takeoff," *Cell*, vol. 74, pp. 773–6, 1993.

[83] D. Mukherjee, A. B. Datta, and P. Chakrabarti, "Crystal Structure of HlyU, the Hemolysin Gene Transcription Activator, from Vibrio cholerae N16961 and Functional Implications," *Biochim. Biophys. Acta*, vol. 1844, pp. 2346–54, 2014.

[84] C. Eicken, M. A. Pennella, X. Chen, K. M. Koshlap, M. L. VanZile, J. C. Sacchettini, and D. P. Giedroc, "A Metal-Ligand-Mediated Intersubunit Allosteric Switch in Related SmtB/ArsR Zinc Sensor Proteins," *J. Mol. Biol.*, vol. 333, no. 4, pp. 683–95, 2003.

[85] D. K. Chakravorty, B. Wang, C. W. Lee, A. J. Guerra, D. P. Giedroc, and K. M. Merz-Jr., "Solution NMR Refinement of a Metal on Bound Protein Using Metal Ion Inclusive Restrained Molecular Dynamics Methods," *J. Biomol. NMR*, vol. 56, pp. 125–37, 2013.

[86] L. Banci, I. Bertini, F. Cantini, S. Ciofi-Baffoni, J. S. cavet, C. Dennison, A. I. Graham, D. R. Harvie, and N. J. Robinson, "NMR Structural Analysis of Cadmium Sensing by Winged Helix Repressor CmtR," *J. Biol. Chem.*, vol. 282, no. 41, pp. 30 181–8, 2007.

[87] K. Nishi, H.-J. Lee, S.-Y. Park, S. J. Baw, S. E. Lee, P. D. Adams, J. H. Rhee, and J.-S. Kim, "Crystal Structure of the Transcriptional Activator HlyU from Vibrio Bulnificus CMCP6," *FEBS Lett.*, vol. 584, pp. 1097–102, 2010.

[88] B. G. Guimaraes, R. L. Barbose, A. S. Soprano, B. M. Campos, T. A. Souza, C. C. C. Tonoli, A. F. P. Leme, M. T. Murakami, and C. E. Benedetti, "Plant pathogenic bacterial utilize biofilm growth-associated repressor (bigr), a novel winged-helix redox switch, to control hydrogen sulphide detoxification under hypoxia," *J. Biol. Chem.*, vol. 286, no. 29, pp. 26 148–57, 2011.

[89] T. Haneishi, A. Terahara, M. Arai, T. Hata, and C. Tamura, "New Antibiotics, Methylenomycins A and B," *J. Antibiot.*, vol. 27, no. 6, pp. 393–9, 1974.

[90] U. Hornemann and D. A. Hopwood, "Isolation and Characterisation of Desepoxy-4, 5-Didehydro-Methylenomycin A. A Precursor of the Antibiotic Methylenomycin A in SCP1+ Strains of Streptomyces coelicolor A3(2)," *Tetrahedron Lett.*, vol. 33, pp. 2977–8, 1978.

[91] L. F. Wright and D. A. Hopwood, "Identification of the Antibiotic Determined by the SCP1 Plasmid of Streptomyces coelicolor A3(2)," *J. Gen. Microbiol.*, vol. 95, pp. 96–106, 1976.

[92] R. Kirkby and D. A. Hopwood, "Genetic Determination of Methylenomycin Synthesis by the SCPl Plasmid of Streptomyces coelicolor A3(2)," *J. Gen. Microbiol.*, vol. 98, pp. 239–52, 1977.

[93] M. Bibb, J. L. Schottel, and S. N. Cohen, "A DNA Cloning System for Interspecies Gene Transfer in Antibiotic-Producing Streptomyces," *Nature*, vol. 284, no. 5756, pp. 526–31, 1980.

[94] K. F. Chater and C. J. Bruton, "Mutational Cloning in Streptomyces and the Isolation of Antibiotic Production Genes," *Gene*, vol. 26, pp. 67–78, 1983.

[95] J. Neal and K. F. Chater, "Nucleotide Sequence Analysis Reveals Similarities Between Proteins Determining Methylenomycin A Resistance in Streptomyces and Tetracycline Resistance in Eubacteria," *Gene*, vol. 58, pp. 229–41, 1987.

[96] K. F. Chater and C. J. Bruton, "Resistance, Regulatory and Production Genes for the Antibiotic Methylenomycin are Clustered," *EMBO J.*, vol. 4, no. 7, pp. 1893–7, 1985.

[97] S. D. Bentley, S. Brown, L. D. Murphy, D. E. Harris, M. A. Quail, J. Parkhill, B. G. Barrell, J. R. McCormick, R. I. Santamaria, R. Losick, M. Yamasaki, H. Kinashi, C. W. Chen, G. Chandra, D. Jakimowicz, H. M. Kieser, T. Kieser, and K. F. Chater, "SCP1, a 356 023 bp Linear Plasmid Adapted to the Ecology and Developmental Biology of its Host, Streptomyces coelicolor A3(2)," *Mol. Microbiol.*, vol. 51, no. 6, pp. 1615–28, 2004.

[98] G. Hobs, A. I. C. Obanye, J. Petty, J. C. Mason, E. Barratt, D. C. J. Gardner, F. Flett, C. P. Smith, P. Broda, and S. G. Oliver, "An Integrated Approach to Studying Regulation of Production of the Antibiotic Methylenomycin by Streptomyces coelicolor A3(2)," *J. Bacteriol.*, vol. 174, no. 5, pp. 1487–94, 1992.

[99] S. O'Rourke, A. Wietzorrek, K. Fowler, C. Corre, G. L. Challis, and K. F. Chater, "Extracellular Signalling, Translational Control, Two Repressors and an Activator all Contribute to the Regulation of Methylenomycin Production in Streptomyces coelicolor," *Mol. Microbiol.*, vol. 71, no. 3, pp. 763–78, 2009.

[100] R. Neal and K. Chater, "Bidirectional Promoter and Terminator Regions Bracket mmr, a Resistance Gene Embedded in the Streptomyces coelicolor A3(2) Gene Cluster Encoding Methylenomycin Production," *Gene*, vol. 100, pp. 75–83, 1991.

[101] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic Local Alignment Search Tool," *J. Mol. Biol.*, vol. 215, pp. 403–10, 1990.

[102] E. Gasteiger, C. H. A. Gattika, S. Davaud, M. R. Wilkins, R. D. Appel, and A. Bairoch, "Protein Identification and Analysis Tools on the ExPASy Server," in *The Proteomics Protocols Handbook*, J. M. Walker, Ed. Humana Press, 2005.

[103] A. Marchler-Bauer and S. H. Bryant, "CD-Search: Protein Domain Annotations on the Fly," *Nucleic Acids Res.*, vol. 32, pp. 327–31, 2004.

[104] A. Marchler-Bauer, J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. DeWeese-Scott, J. H. Fong, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, S. He, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, C. A. Liebert, C. Liu, F. Lu, S. Lu, G. H. Marchler, M. Mullokandov, J. S. Song, A. Tasneem, N. Thanki, R. A. Yamashita, D. Zhang, N. Zhang, and S. H. Bryant, "CDD: Specific Functional Annotation with the Conserved Domain Database," *Nucleic Acids Res.*, vol. 37, pp. 205–10, 2009.

[105] A. Marchler-Bauer, S. Lu, J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. DeWeese-Scott, J. H. Fong, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, F. Lu, G. H. Marchler, M. Mullokandov, M. V. Omelchenko, C. L. Robertson, J. S. Song, N. Thanki, R. A. Yamashita, D. Zhang, N. Zhang, C. Zheng, and S. H. Bryant, "CDD: a Conserved Domain Database for the Functional Annotation of Proteins," *Nucleic Acids Res.*, vol. 39, pp. 225–9, 2011.

[106] P. Stothard, G. van Domselaar, S. S. S, A. Guo, B. O'Neill, J. Cruz, M. Ellison, and D. S. W. DS, "BacMap: An Interactive Picture Atlas of Annotated Bacterial Genomes," *Nucleic Acids Res.*, vol. 33, pp. D317–20, 2005.

[107] E. D. Castro, C. J. A. Sigrist, A. Gattiker, V. Bulliard, P. S. Langendijk-Genevaux, E. Gasteiger, A. Bairoch, and N. Hulo, "ScanProsite: Detection of PROSITE Signature Matches and ProRule-Associated Functional and Structural Residues in Proteins," *Nucleic Acids Res.*, vol. 34, pp. W362–5, 2006.

[108] C. J. A. Sigrist, E. D. Castro, L. Cerutti, B. A. Cuche, N. Hulo, A. Bridge, L. Bougueleret, and I. Xenarios, "New and Continuing Developments at PROSITE," *Nucleic Acids Res.*, vol. 41, pp. D344–7, 2013.

[109] C. J. A. Sigrist, L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, and P. Bucher, "PROSITE: A Documented Database Using Patterns and Profiles as Motif Descriptors," *Brief. Bioinform.*, vol. 3, no. 3, pp. 265–74, 2002.

[110] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, "GenBank," *Nucleic Acids Res.*, vol. 41, pp. D36–42, 2013.

[111] T. L. Bailey and C. Elkan, "Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Biopolymers," in *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology.* AAAI Press, 1994, pp. 28–36.

[112] L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, and M. J. E. Sternberg, "The Phyre2 Web Portal for Protein Modelling, Prediction and Analysis," *Nat. Protoc.*, vol. 10, pp. 845–58, 2015.

[113] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhangan, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs," *Nucleic Acids Res.*, vol. 25, pp. 3389–402, 1997.

[114] Y. Zhang and J. Skolnick, "Scoring function for Automated Assessment of Protein Structure Template Quality," *Proteins.*, vol. 57, no. 4, pp. 702–10, 2004.

[115] C. Dominguez, R. Boelens, and A. M. Bonvin, "HADDOCK: A Protein-Protein Docking Approach Based on Biochemical and/or Biophysical Information," *J. Am. Chem. Soc.*, vol. 125, pp. 1731–7, 2003.

[116] S. de Vries, A. van Dijk, M. Krzeminski, M. van Dijk, A. Thureau, V. Hsu, T. Wassenaar, and A. Bonvin, "HADDOCK versus HADDOCK: New Features and Performance of HADDOCK2.0 on the CAPRI Targets," *Proteins.*, vol. 69, pp. 726–33, 2007.

[117] C. W. Lee, D. K. Chakravorty, F.-M. J. Chang, H. Reyes-Caballero, Y. Ye, K. M. Merz, and D. P. Giedroc, "Solution Structure of Mycobacterium tuberculosis NmtR in the Apo State: Insights into Ni(II)-Mediated Allostery," *Biochemistry*, vol. 51, no. 12, pp. 2619–29, 2012.

[118] A. Dong, X. Xu, H. Zheng, A. M. Edwards, A. Joachimiak, and A. Savchenko, 2008, unpublished.

[119] J. R. Doroghazi, J. C. Albright, A. W. Goering, K.-S. Ju, R. R. Haines, K. A. Tchalukov, D. P. Labeda, N. L. Kelleher, and W. W. Metcalf, "A Roadmap for Natural Product Discovery Based on Large-Scale Genomics and Metabolomics," 2014, unpublished, NCBI GI:664208801.

[120] I. J. Dang, M. Huntemann, J. Han, A. Chen, N. Kyrpides, K. Mavromatis, V. Markowitz, K. Palaniappan, N. Ivanova, A. Schaumberg, A. Pati, K. Liolios, H. P. Nordberg, M. N. Cantor, S. X. Hua, and T. Woyke, 2013, unpublished, NCBI GI: 487392204.

[121] G. H. Gonnet, M. A. Cohen, and S. H. Benner, "Exhaustive Matching of the Entire Protein Sequence Database," *Science*, vol. 256, pp. 1443–5, 1992.

[122] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins, "Clustal W and Clustal X version 2.0," *Bioinformatics*, vol. 23, no. 21, pp. 2947–8, 2007.

[123] C. Park, J. L. Campbell, and W. A. Goddard-III, "Design Superiority of Palindromic DNA Sites for Site-Specific Recognition of Proteins: Tests Using Protein Stitchery," *Proc. Nat. Acad. Sci. USE*, vol. 90, pp. 4892–6, 1993.

[124] *Champion pET Directional TOPO Expression Kits*, invitrogen, June 2007.

[125] F. Studier and B. A. Moffat, "Use of Bacteriophage T7 RNA Polymerase to Direct Selective High-Level Expression of Cloned Genes," *J. Mol. Biol.*, vol. 189, pp. 113–30, 1986.

[126] F. Cedrone, S. Niel, S. Roca, T. Bhatnagar, N. Ait-Abdelkader, C. Torre, H. Krumm, A. Maichele, M. T. Reetz, and J. C. Baratti, "Directed Evolution of the Epoxide Hydrolase from Aspergillus niger," *Biocatal. Biotransfor.*, vol. 21, no. 6, pp. 357–64, 2003.

[127] F. Wang, Y. Min, and X. Geng, "Fast Separations of Intact Proteins by Liquid Chromatography," *J. Sep. Sci.*, vol. 35, no. 22, pp. 3033–45, 2012.

[128] J. Giacometti and D. Josic, "Protein and Peptide Separations," in *Liquid Chromatography Applications*, S. Fanali, P. R. Haddad, C. F. Poole, P. Schoenmakers, and D. Lloyd, Eds. Elsevier, 2013, ch. 7, pp. 149–84.

[129] B. Goke and V. Keim, "HPLC and FPLC - Recent Progress in the Use of Automated Chromatography Systems for Resolution of Pancreatic Secretory Proteins," *Int. J. Pancreatol.*, vol. 11, no. 2, pp. 109–16, 1992.

[130] *Strategies for Protein Purification Handbook*, GE Healthcare, 2010.

[131] D.S.Hage, J. A. Anguizola, R. Li, R. Matsuda, E. Papastavros, E. Pfaunmiller, M. Sobansky, and X. Zheng, "Affinity Chromatography," in *Liquid Chromatography Applications*, S. Fanali, P. R. Haddad, C. F. Poole, P. Schoenmakers, and D. Lloyd, Eds. Elsevier, 2013, ch. 1, pp. 1–24.

[132] J. Porath, J. Carlsson, I. Olsson, and B. Belfrage, "Meal Chelate Affinity Chromatography: A New Approach to Protein Fraction," *Nature*, vol. 258, pp. 598–9, 1975.

[133] R. Guitierrez, E. M. M. del Valle, and M. A. Galan, "Immobilised Metal-Ion Affinity Chromatography: Status and Trends," *Sep. Purif. Rev.*, vol. 36, pp. 71–111, 2007.

[134] A. Schwarz, "Affinity Purification of Monoclonal Antibodies," in *Affinity Chromatography Methods and Protocols*, ser. Methods in Molecular Biology, P. Bailon, G. K. Ehrlich, W.-J. Fung, and W. Berthold, Eds. Humana Press, 2000, vol. 147, ch. 5, pp. 49–56.

[135] J. Thorner, S. D. Emr, and J. N. Abelson, Eds., *Applications of Chimeric Genes and Hybrid Proteins, Part A: Gene Expression and Protein Purification*, ser. Methods in Enzymology. Sandiego, California: Academic Press, 2000, vol. 326.

[136] C. B. Parsy, C. J. Chapman, A. C. Barnes, J. F. Robertson, and A. Murray, "Two-Step Method to Isolate Target Recombinant Protein from Co-Purified Bacterial Contaminant SlyD After Immobilised Metal Affinity Chromatography," *J. Chromatogr. B*, vol. 853, no. 1, pp. 314–9, 2007.

[137] *HisTrap HP, 1 mL and 5 mL*, GE Healthcare, 2009.

[138] J.-D. Pedelacq, S. Cabantous, T. Tran, T. C. Terwilliger, and G. S. Waldo, "Engineering and Characterization of a Superfolder Green Fluorescent Protein," *Nat. Biotechnol.*, vol. 24, no. 1, pp. 79–88, 2006.

[139] X. Wu, D. Wu, Z. Lu, W. Chen, X. Hu, and Y. Ding, "A Novel Method for High-Level Production of TEV Protease by Superfolder GFP Tag," *J. Biomed. Biotechnol.*, vol. 2009, 2009.

[140] J. E. Tropea, S. Cherry, and D. S. Waugh, "Expression and Purification of Soluble His6-Tagged TEV Protease," in *Methods in Molecular Biology: High Throughput Protein Expression and Purification*, S. A. Doyle, Ed. Humana Press, 2009, vol. 498.

[141] J. B. Fenn, M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse, "Electrospray Ionization for Mass Spectrometry of Large Biomolecules," *Science*, vol. 245, no. 4926, pp. 64–71, 1989.

[142] B. Nordén, A. Rodger, and T. Dafforn, *Linear Dichroism and Circular Dichroism.* RSC Publishing, 2010.

[143] N. Greenfield and G. D. Fasman, "Computed Circular Dichroism Spectra for the Evaluation of Protein Conformation," *Biochemistry*, vol. 8, no. 10, pp. 4108–16, 1969.

[144] T. Creighton, Ed., *Encyclopedia of Molecular Biology.* John Wiley and Sons, 1999, vol. 1.

[145] N. Greenfield, "Using Circular Dichroism Spectra to Estimate Protein Secondary Structure," *Nat. Protoc.*, vol. 1, no. 6, pp. 2876–90, 2007.

[146] S. M. Kelly, T. J. Jess, and N. C. Price, "How to Study Proteins by Circular Dichroism," *Biochim. Biophys. Acta*, vol. 1751, pp. 119–39, 2005.

[147] C. Ó'Fágáin, P. M. Cummins, and B. F. O'Connor, "Gel-Filtration Chromatography," in *Protein Chromatography*, D. Walls and S. T. Loughran, Eds. Humana Press, 2011, vol. 681, pp. 25–33.

[148] Uncertainty Calculator. [Online]. Available: http://web.mst.edu/~gbert/JAVA/uncertainty.HTML

[149] G. Bertrand. Uncertainties in Chemical Calculations. [Online]. Available: http://web.mst.edu/~gbert/STATS.pdf

[150] D. Moinier, D. Slyemi, D. Byrne, S. Lignon, R. Lebrun, E. Talla, and V. Bonnefoya, "An ArsR/SmtB Family Member Is Involved in the Regulation by Arsenic of the Arsenite Oxidase Operon in Thiomonas arsenitoxydans," *Appl. Environ. Microb.*, vol. 80, no. 20, pp. 6413–26, 2014.

[151] G. B. Ralston, *Introduction to Analytical Ultracentrifugation.* Beckman, 1993.

[152] P. Schuck, "Size Distribution Analysis of Macromolecules by Sedimentation Velocity Ultracentrifugation and Lamm Equation Modeling," *Biophys. J.*, vol. 78, pp. 1606–19, 2000.

[153] C. J. Roberts, "Non-Native Protein Aggregation Kinetics," *Biotechnol. Bioeng.*, vol. 98, no. 5, pp. 927–38, 2007.

[154] M. D. Biggins, "Animal Transcription Networks as Highly Connected, Quantitative Continua," *Dev. Cell*, vol. 21, no. 4, pp. 611–26, 2011.

[155] (2015) Clone Manager 9 for Windows. [Online]. Available: http://www.scied.com/pr_cmbas.htm

[156] M. M. Garner and A. Revzin, "A Gel Electrophoresis Method for Quantifying the Binding of Proteins to Specific DNA Regions: Application to Components of the Escherichia coli Lactose Operon Regulatory System," *Nucleic Acids Res.*, vol. 9, no. 13, pp. 3047–60, 1981.

[157] M. Fried and D. M. Crothers, "Equilibria and Kinetics of lac Repressor-Operator Interactions by Polyacrylamide Gel Electrophoresis," *Nucleic Acids Res.*, vol. 9, no. 23, pp. 6505–25, 1981.

[158] H. Rilbe, "Basic Theory of Electrophoresis: Definitions, Terminology and Comparison of the Basic Techniques," in *Electrophoretic Techniqes*, C. F. Simpson and M. Whittaker, Eds. Academic Press, 1983, pp. 1–26.

[159] (2015) DIG DNA Labelling and Detection Kit — Sigma-Aldrich. [Online]. Available: http://www.sigmaaldrich.com/catalog/product/roche/11093657910?lang=en&region=GB

[160] A. Chant, C. M. Kraemer-Pecore, R. Watkin, and G. G. Kneale, "Attachment of a Histidine Tag to the Minimal Zinc Finger Protein of the Aspergillus nidulans Gene Regulatory Protein AreA Causes a Conformational Change at the DNA-Binding Site," *Protein Expr. Purif.*, vol. 39, no. 2, pp. 152–9, 2005.

[161] B. C. Stanton, A. A. K. Nielsen, A. Tamsir, K. Clancy, T. Peterson, and C. A. Voigt, "Genomic Mining of Prokaryotic Repressors for Orthogonal Logic Gates," *Nat. Chem. Biol.*, vol. 10, pp. 99–105, 2014.

[162] C. Corre, L. Song, M. E. Whitehead, R. J. Deeth, and G. L. Challis, "Biosynthesis of the Exomethylene and Epoxide Functions of the Methylenomycin Antibiotics in Streptomyces coelicolor A3(2)," unpublished.

[163] J. Jernow, W. Tautz, P. Rosen, and T. H. Williams, "Methylenomycin B: Revised Structure and Total Syntheses," *J. Org Chem.*, vol. 44, no. 23, pp. 4212–3, 1979.

[164] A. Shatz, E. Bugle, and S. A. Waksman, "Streptomycin, a Substance Exhibiting Antibiotic Activity Against Gram-Positive and Gram-Negative Bacteria," *Exp. Biol. Med.*, vol. 55, no. 1, pp. 66–9, 1944.

[165] C. Corre, L. Song, S. O'Rourke, K. F. Chater, and G. L. Challis, "2-Alkyl-4-Hydroxymethylfuran-3-Carboxylic Acids, Antibiotic Production Inducers Discovered by Streptomyces coelicolor Genome Mining," *Proc. Nat. Acad. Sci. USE*, vol. 104, no. 45, pp. 17 510–5, 2008.

[166] P. J. Harrison, N. Malet, D. Rea, S. Zhou, K. Styles, V. Fulop, G. L. Challis, and C. Corre, "Structure-Function Studies of MmfR, a Transcriptional Repressor That Prevent Antibiotic Biosynthesis in Streptomyces coelicolor A3(2)," unpublished.

[167] M. Becker-Andre and K. Hahlbrock, "Absolute mRNA Quantification Using the Polymerase Chain Reaction (PCR). A Novel Approach by a PCR Aided Transcript Titration Assay (PATTY)," *Nucleic Acids Res.*, vol. 17, no. 22, pp. 9437–46, 1989.

[168] G. Gilliland, S. Perrin, K. Blanchard, and H. F. Bunn, "Analysis of Cytokine mRNA and DNA: Detection and Quantitation by Competitive Polymerase Chain Reaction," *Proc. Nat. Acad. Sci. USE*, vol. 87, pp. 2725–9, 1990.

[169] W. M. Freeman, S. J. Walker, and K. E. Vrana, "Quantitative RT-PCR: Pitfalls and Potential," *Biotechniques*, vol. 26, no. 1, pp. 112–25, 1999.

[170] A. Craney, T. Hohenauer, Y. Xu, N. K. Navani, Y. Li, and J. Nodwell, "A Synthetic luxCDABE Gene Cluster Optimized for Expression in High-GC Bacteria," *Nucleic Acids Res.*, vol. 35, no. 6, p. e46, 2007.

[171] (2015) Quick Ligation Protocol (M2200) — NEB. [Online]. Available: https://www.neb.com/protocols/1/01/01/quick-ligation-protocol

[172] (2015) IP-Free Synthetic Chromogenic and Fluorescent Protein Vectors - DNA2.0. [Online]. Available: https://www.dna20.com/products/protein-paintbox?exp=3

[173] L. Whitmore and B. A. Wallace, "Protein Secondary Structure Analyses from Circular Dichroism Spectroscopy: Methods and Reference Databases," *Biopolymers*, vol. 89, pp. 392–400, 2008.

[174] L. Whitmore and B. A. Wallace, "DICHROWEB: An Online Server for Protein Secondary Structure Analyses from Circular Dichroism Spectroscopic Data," *Nucleic Acids Res.*, vol. 32, pp. W668–73, 2004.

[175] A. Lobley, L. Whitmore, and B. A. Wallace, "DICHROWEB: An Interactive Website for the Analysis of Protein Secondary Structure from Circular Dichroism Spectra," *Bioinformatics*, vol. 18, pp. 211–2, 2002.

[176] N. Sreerema and R. W. Woody, "A Self-Consistent Method for the Analysis of Protein Secondary Structure from Circular Dichroism," *Anal. Biochem*, vol. 208, pp. 32–44, 1993.

[177] N. Sreerema, S. Y. Venyaminov, and R. W. Woody, "Estimation of the Number of Helical and Strand Segments in Proteins using CD Spectroscopy," *Protein Sci.*, vol. 8, pp. 370–80, 1999.

[178] S. W. Provencher and J. Glockner, "Estimation of Globular Protein Secondary Structure from Circular Dichroism," *Biochemistry*, vol. 20, pp. 33–7, 1981.

[179] I. H. M. van Stokkum, H. J. W. Spoelder, M. Bloemendal, R. van Grondelle, and F. C. A. Groen, "Estimation of Protein Secondary Structure and Error Analysis from CD Spectra," *Anal. Biochem*, vol. 191, pp. 110–8, 1990.

[180] L. A. Compton and W. C. Johnson, "Analysis of Protein Circular Dichroism Spectra for Secondary Structure Using a Simple Matrix Multiplication," *Anal. Biochem*, vol. 155, pp. 155–67, 1986.

[181] P. Manavalan and W. C. Johnson, "Variable Selection Method Improves the Prediction of Protein Secondary Structure from Circular Dichroism Spectra," *Anal. Biochem*, vol. 167, pp. 76–85, 1987.

[182] N. Sreerema and R. W. Woody, "Estimation of Protein Secondary Structure from Circular Dichroism Spectra: Comparison of CONTIN, SELCON, and CDSSTR Methods with an Expanded Reference Set," *Anal. Biochem*, vol. 287, no. 2, pp. 252–60, 2000.

[183] T. Morii, C. S. Lim, and S. N. Mukherjee, *The Physics of the Standard Model and Beyond*. Singapore: World Scientific, 2004.

[184] M. H. Levitt, *Spin Dynamics: Basics of Nuclear Magnetic Resonance*. Chichester, England: Wiley, 2001.

[185] J. Marley, M. Lu, and C. Bracken, "A Method for Efficient Isotopic Labeling of Recombinant Proteins," *J. Biomol. NMR*, vol. 20, no. 1, pp. 71–5, 2001.

[186] F. A. Hopf, R. F. Shea, and M. O. Scully, "Theory of Optical Free-Induction Decay and Two-Photon Superradiance," *Phys. Rev. A*, vol. 7, no. 6, pp. 2105–10, 1973.

[187] G. Bodenhausen and D. J. Ruben, "Natural Abundance Nitrogen-15 NMR by Enhance Heteronuclear Spectroscopy," *Chem. Phys. Lett.*, vol. 69, no. 1, pp. 185–9, 1980.

[188] M. A. L. Eriksson, T. Hard, and L. Nilsson, "On the pH Dependence of Amide Proton Exchange Rates in Proteins," *Biophys. J.*, vol. 69, no. 2, pp. 329–39, 1995.

[189] A. P. Golovanov, G. M. Hautbergue, S. A. Wilson, and L.-Y. Lian, "A Simple Method for Improving Protein Solubility and Long-Term Stability," *J. Am. Chem. Soc.*, vol. 126, pp. 8933–9, 2004.

[190] L. P. McIntosh and F. W. Dahlquist, "Biosynthetic Incorporation of 15N and 13C for Assignment and Interpretation of Nuclear Magnetic Resonance Spectra of Proteins," *Quart. Rev. Biophys*, vol. 23, no. 1, pp. 1–38, 1990.

[191] J. Schaefer and E. O. Stejskal, "Carbon-13 Nuclear Magnetic Resonance of Polymers Spinning at the Magic Angle," *J. Am. Chem. Soc.*, vol. 98, no. 4, pp. 1031–2, 1976.

[192] I. Bertini, C. Luchinat, G. Parigi, E. Ravera, B. Reif, and P. Turano, "Solid-state NMR of Proteins Sedimented by Ultracentrifugation," *Proc. Nat. Acad. Sci. USE*, vol. 108, no. 26, pp. 10 396–9, 2011.

[193] A. Pines, M. G. Gibby, and J. S. Waugh, "Proton-Enhanced NMR of Dilute Spins in Solids," *J. Chem. Phys.*, vol. 59, pp. 569–90, 1973.

[194] D. Brewster, *A Treatise on Optics*. London: Longman, Rees, Orme, Brown and Green, 1831.

[195] T. Young, "Lecture 39," in *A Course of Lectures on Natural Philosophy and the Mechanical Arts*. Bury: William Savage, 1807, vol. 1.

[196] W. H. Bragg and W. L. Bragg, "The Reflexion of X-rays by Crystals," *Proc. R. Soc. Lond.*, vol. 88, no. 605, pp. 428–38, 1913.

[197] T. L. Blundell and L. N. Johnson, *Protein Crystallography*, ser. Molecular Biology. Academic Press, 1976.

[198] P. P. Ewald, "VII. Das reziproke Gitter "in der Strukturtheorie," *Z. Kristollagr. Miner.*, vol. 56, no. 129, 1921.

[199] P. P. Ewald, "Introduction to the Dynamical Theory of X-ray Diffraction," *Acta Crystallogr. A*, vol. 25, no. 1, 1969.

[200] D. Sherwood and J. Cooper, *Crystals, X-rays and Proteins: Comprehensive Protein Crystallography*. Oxford, UK: Oxford Press, 2011.

[201] A. Tulinski, "The Protein Structure Project, 1950-1959: First Concerted Effort Of a Protein Structure Determination In the U.S," *Annu. Rep. Med Chem.*, vol. 31, pp. 357–60, 1999.

[202] F. Gorrec, "The MORPHEUS Protein Crystallization Screen," *J. Appl. Crystallogr.*, vol. 42, pp. 1035–42, 2009.

[203] S. Radaev, S. Li, and P. D. Sun, "A Survey of Protein-Protein Complex Crystallizations," *Acta Crystallogr. D*, vol. 62, pp. 605–12, 2006.

[204] J. Newman, D. Egan, T. S. Walter, R. Meged, I. Berry, M. B. Jelloul, J. L. Sussman, D. I. Stuart, and A. Perrakis, "Towards Rationalization of Crystallization Screening for Small- to Medium-Sized Academic Laboratories: the PACT/JCSG+ Strategy," *Acta Crystallogr. D*, vol. 61, pp. 1426–31, 2005.

[205] N. E. Chayen, "Turning Protein Crystallisation from an Art into a Science," *Curr. Opin. Struc. Biol.*, vol. 14, no. 5, pp. 577–83, 2004.

[206] D. B. Cowie and G. N. Cohen, "Biosynthesis by E. coli of Active Proteins Containing Selenium Instead of Sulphur," *Biochim. Biophys. Acta*, vol. 26, pp. 252–61, 1975.

[207] W. A. Hendrickson, J. R. Horton, and D. M. LeMaster, "Selenomethionyl Proteins Produced for Analysis by Multiwavelength Anomalous Diffraction (MAD): a Vehicle for Direct Determination of Three-Dimensional Structure," *EMBO J.*, vol. 9, pp. 1665–72, 1990.

[208] W. A. Hendrickson, "Determination of Macromolecular Structures from Anomalous Diffraction of Synchrotron Radiation," *Science*, vol. 254, pp. 51–8, 1991.

[209] *GeneJET Gel Extraction Kit*, 12th ed., Thermo Scientific, Carlsbad, California, 2015.

[210] *GeneJET Plasmid Miniprep Kit, K0502, K0503*, Thermo Scientific, 2013.

[211] *Rneasy Plus Micro Handbook*, QIAGEN, Austin, Texas, July 2007.

[212] *TURBO DNA-free Kit: TURBO DNase Treatment and Removal Reagents*, G ed., ambion, Carlsbad, California, 2012.

[213] *SuperScript II Reverse Transcriptase*, invitrogen, Carlsbad, California, May 2010.

[214] *QIAquick Spin Handbook*, QIAGEN, Austin, Texas, April 2015.

[215] *QuikChange Lightning Site-Directed Mutagenesis Kit*, E.01 ed., Agilent Technologies, 2013.

[216] A. Sivashanmugam, V. Murray, C. Cui, Y. Zhang, J. Wang, and Q. Li, "Practical Protocols for Production of Very High Yields of Recombinant Proteins Using Escherichia coli," *Protein Sci.*, vol. 18, no. 5, pp. 936–48, 2009.

[217] QuikChange Primer Design — Agilent Technologies. [Online]. Available: http://www.genomics.agilent.com/primerDesignProgram.jsp

[218] Sedfit Help Web. [Online]. Available: http://www.analyticalultracentrifugation.com/default.htm

[219] (2012, August) Sednterp Main Page. [Online]. Available: http://bitcwiki.sr.unh.edu/index.php/Main_Page

[220] *FastDNA SPIN Kit for Soil*, MP Biomedicals, Santa Ana, California.